Transactions of the I·R·E

Professional Group on INFORMATION THEORY

PGIT-4 SEPTEMBER 1954

1954 SYMPOSIUM ON INFORMATION THEORY

held at
Massachusetts Institute of Technology
Cambridge, Massachusetts

September 15-17, 1954

The Institute of Radio Engineers

IRE PROFESSIONAL GROUP ON INFORMATION THEORY

The Professional Group on Information Theory is an organization, within the framework of the IRE, of members with principal professional interest in Information Theory. All members of the IRE are eligible for membership in the Group and will receive all Group publications upon payment of prescribed assessments.

Annual Assessment: $2.00

Administrative Committee

Chairman: William G. Tuller, Melpar, Inc., Alexandria, Va.

Vice Chairman: Louis A. deRosa, Federal Telecommunications Laboratories, Inc., Nutley, N. J.

Secretary: Harold R. Holloway, Sylvania Electric Products, Inc., Bayside, L. I., N. Y.

R. M. Fano
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge 39, Massachusetts

M. J. E. Golay
Squier Signal Laboratory
Fort Monmouth, New Jersey

C. H. Page
National Bureau of Standards
Washington 25, D. C.

M. J. DiToro
Fairchild Guided Missile Laboratory
Wyandanch, Long Island, New York

Meyer Leifer
Electronic Defense Laboratory
Syvania Electronic Products, Inc.
Mountain View, P.O. Box 205, Calif.

W. D. White
Airborne Instruments Laboratory, Inc.
160 Old Country Road
Mineola, New York

Nathan Marchand
Marchand Electronics
Greenwich, Connecticut

Winslow Palmer
Sperry Gyroscope Company
Great Neck, New York

Wilbur B. Davenport, Jr.
Lincoln Laboratories
Massachusetts Institute of Technology
Cambridge 39, Massachusetts

Laurin G. Fischer
Federal Telecommunications Labs., Inc.
Nutley, New Jersey

Ernest R. Kretzmer
Bell Telephone Laboratories
Murray Hill, New Jersey

Bernard M. Oliver
Hewlett-Packard Corporation
Palo-Alto, California

TRANSACTIONS OF THE I•R•E®
Professional Group on Information Theory

Published by the Institute of Radio Engineers, Inc., for the Professional Group on Information Theory at 1 East 79th Street, New York 21, N. Y. Responsibility for the contents rests upon the authors, and not upon the Institute, the Group or its members. Individual copies available for sale to IRE-PGIT members at $3.35; to IRE members at $5.00; and to non-members at $10.00.

TRANSACTIONS

of the

1954 SYMPOSIUM ON INFORMATION THEORY

held at

Massachusetts Institute of Technology, Cambridge, Massachusetts

September 15-17, 1954

Organized by

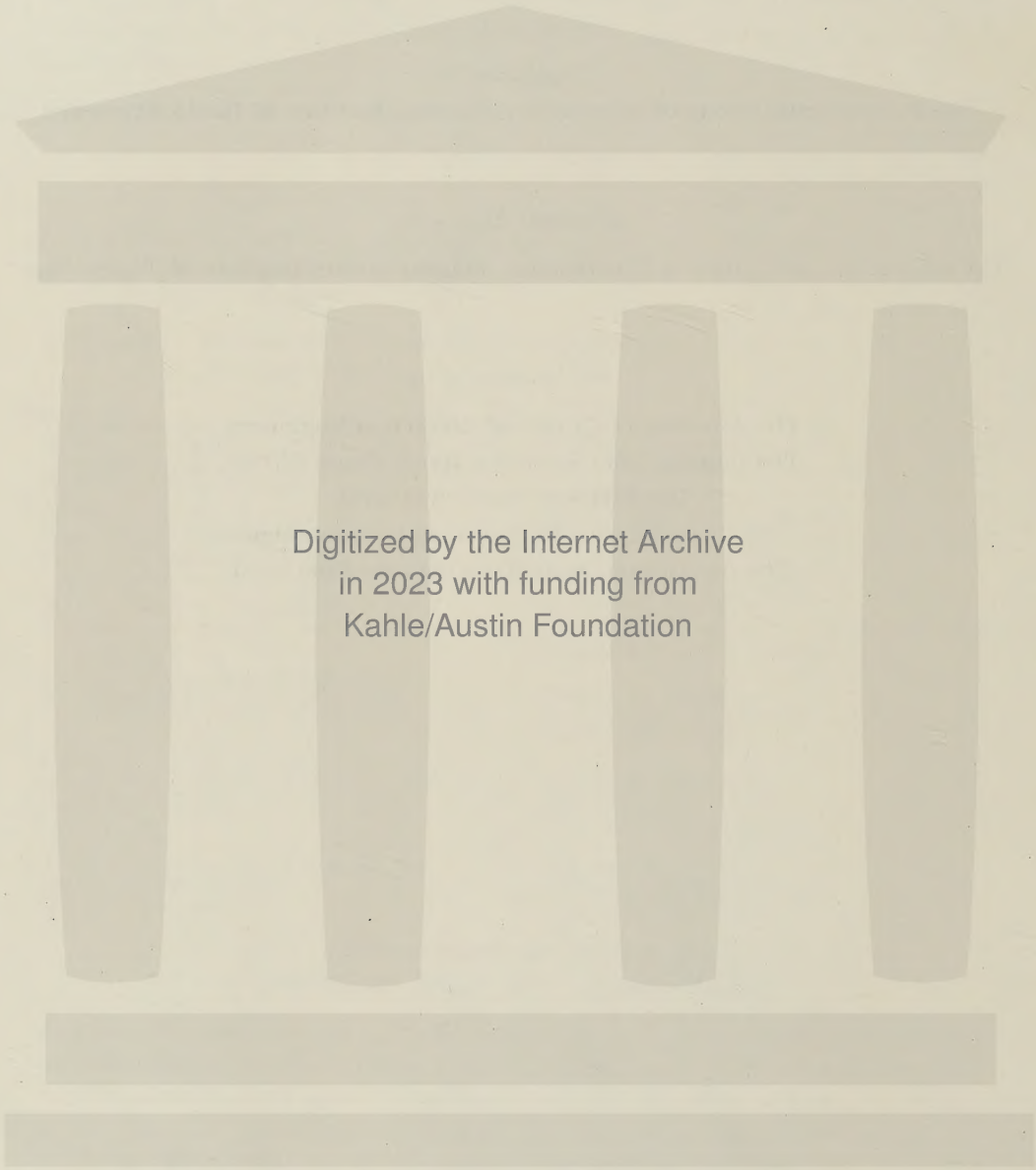
The Professional Group on Information Theory, Institute of Radio Engineers

In cooperation with

The Research Laboratory of Electronics, Massachusetts Institute of Technology

and sponsored by

The American Institute of Electrical Engineers
The International Scientific Radio Union (URSI)
The Office of Naval Research
The Signal Corps Engineering Laboratories
The Air Research and Development Command



CONTENTS AND ABSTRACTS

PREFACE

The 1954 Symposium on Information Theory is the regular yearly symposium organized by the Professional Group on Information Theory and held alternately on the East and West Coasts.

In view of the widespread interest in information theory, the PGIT has invited other organizations to join in the planning and sponsoring of the symposium so as to make it the one outstanding occasion for presentation and discussion of the most recent, significant advances made in the field.

We are most happy to list among the sponsors of this symposium our sister organization, the American Institute of Electrical Engineers, and the International Scientific Radio Union (URSI). We are particularly thankful to the Research Laboratory of Electronics of the Massachusetts Institute of Technology for serving as host to the Symposium, and to the Office of Naval Research, the Air Research and Development Command, and the Signal Corps Engineering Laboratories for the financial support provided by them through their joint contract with the Research Laboratory of Electronics.

The TRANSACTIONS are published prior to the Symposium to allow the participants to familiarize themselves with the papers. It is hoped that the spirited discussion that will undoubtedly arise from a better-informed audience will generate new ideas and stimulate future research.

Meeting the schedule required to publish these TRANSACTIONS has subjected many people to pressure and inconveniences. We are very grateful to the individual authors for meeting unusually stringent deadlines; to the organizing committee for the careful planning and review of papers; and to the staff of IRE Headquarters for the prompt publication and distribution of these TRANSACTIONS.

W. G. Tuller, Chairman
Professional Group on Information Theory
Institute of Radio Engineers

A NEW BASIC THEOREM OF INFORMATION THEORY *

Amiel Feinstein
Research Laboratory of Electronics, Massachusetts Institute of Technology
Cambridge, Massachusetts

INTRODUCTION

Information theory, in the restricted sense used in this paper, originated in the classical paper of C. E. Shannon, in which he gave a precise mathematical definition for the intuitive notion of information. In terms of this definition it was possible to define precisely the notion of a communication channel and its capacity. Like all definitions that purport to deal with intuitive concepts, the reasonability and usefulness of these definitions depend for the most part on theorems whose hypotheses are given in terms of the new definitions but whose conclusions are in terms of previously defined concepts. The theorems in question are called the fundamental theorems for noiseless and noisy channels. We shall deal exclusively with noisy channels.

By a communication channel we mean, in simplest terms, an apparatus for signaling from one point to another. The abstracted properties of a channel that will concern us are: (a) a finite set of signals that may be transmitted; (b) a set (not necessarily finite) of signals that may be received; (c) the probability (or probability density) of the reception of any particular signal when the signal transmitted is specified. A simple telegraph system is a concrete example. The transmitted signals are a long pulse, a short pulse, and a pause. If there is no noise in the wire, the possible received signals are identical with the transmitted signals. If there is noise in the wire, the received signals will be mutilations of the transmitted signals, and the conditional probability will depend on the statistical characteristics of the noise present.

We shall now sketch the definitions and theorems mentioned. Let X be a finite abstract set of elements x, and let p( ) be a probability distribution on X. We define the "information content" of X by the expression $-\sum_X p(x) \log_2 p(x)$, where the base 2 simply determines the fundamental unit of information, called "bit". One intuitive way of looking at this definition is to consider a machine that picks, in a purely random way but with the given probabilities, one x per second from X. Then $-\log_2 p(x_o)$ may be considered as the information or surprise associated with the event that x_o actually came up. If each event x consists of several events, that is, if $x = \{a, b, \ldots\}$, we have the following meaningful result: $H(X) \leq H(A) + H(B) + \ldots$ with equality if, and only if, the events a, b, ... are mutually independent.

We are now in a position to discuss the first fundamental theorem. We set ourselves the following situation. We have the set X. Suppose, further, that we have some "alphabet" of D "letters" which we may take as $0, \ldots, D-1$. We wish to associate to each x a sequence of integers $0, \ldots, D-1$ in such a way that no sequence shall be an extension of some shorter sequence (for otherwise they would be distinguishable by virtue of their length, which amounts to introducing a $D+1^{th}$ "variable"). Now it is easy to show that a set of D elements has a maximum information content when each element has the same probability, namely 1/D. Suppose now that with each x we associate a sequence of length N_x. The maximum amount of information obtainable by "specifying" that sequence is $N_x \log_2 D$ bits. Suppose $N_x \log_2 D = -\log_2 p(x)$; then $\sum_X p(x) N_x = H(X)/\log_2 D$ is the average length of the sequence. The first

* This work was supported in part by the Signal Corps; the Office of Scientific Research, Air Research and Development Command; and the Office of Naval Research.

fundamental theorem now states that if we content ourselves with representing sequences of x's by sequences of integers $0, \ldots, D-1$, then if we choose our x-sequences sufficiently long, the sequences of integers representing them will have an average length as little greater then $H(X)/\log_2 D$ as desired, but that it is not possible to do any better than this.

To discuss the second fundamental theorem, we now take, as usual, X to be the set of transmitted messages and Y the set of received signals. For simplicity we take Y finite. The conditional probability mentioned above we denote by $p(y/x)$. Let $p(\ )$ be a probability distribution over X, whose meaning is the probability of each x being transmitted. Then the average amount of information being fed into the channel is $H(X) = -\sum_X p(x) \log_2 p(x)$. Since in general the reception of a y does not uniquely specify the x transmitted, we inquire how much information was lost in transmission. To determine this, we note that, inasmuch as the x was completely specified at the time of transmission, the amount of information lost is simply the amount of information necessary (on the average, of course) to specify the x. Having received y, our knowledge of the respective probability of each x having been the one transmitted is given by $p(x/y)$. The average information needed to specify x is now $-\sum_X p(x/y) \log_2 p(x/y)$. We must now average this expression over the set of possible y's. We obtain finally

$$\sum_Y p(y) \left[-\sum_X p(x/y) \log_2 p(x/y) \right] = -\sum_Y \sum_X p(x, y) \log_2 p(x/y) \equiv H(X/Y)$$

often called the equivocation of the channel. The rate at which information is received through the channel is therefore $R = H(X) - H(X/Y)$. A precise statement of the fundamental theorem for noisy channels is given in section II.

I. For the sake of definiteness we begin by stating a few definitions and subsequent lemmas, more or less familiar.

Let X and Y be abstract sets consisting of a finite number, α and β, of points x and y. Let $p(\ )$ be a probability distribution over X, and for each $x \in X$ let $p(\ /x)$ denote a probability distribution over Y. The totality of objects thus far defined will be called a communication channel.

The situation envisaged is that X represents a set of symbols to be transmitted and Y represents the set of possible received signals. Then $p(x)$ is the a priori probability of the transmission of a given symbol x, and $p(R/x)$ is the probability of the received signal lying in a subset R of Y, given that x has been transmitted. Clearly, $\sum_{x \in Q} p(x)\, p(R/x)$ represents the joint probability of R and a subset Q of X, and will be written as $p(Q, R)$. Further, $p(X, R) \equiv p(R)$ represents the absolute probability of the received signal lying in R. (The use of p for various different probabilities should not cause any confusion.)

The "information rate" of the channel "source" X is defined by $H(X) = -\sum_X p(x) \log p(x)$, where here and in the future the base of the logarithm is 2. The "reception rate" of the channel is defined by the expression

$$\sum_X \sum_Y p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)} \geq 0$$

If we define the "equivocation" $H(X/Y) = -\sum_X \sum_Y p(x,y) \log p(x/y)$ then the reception rate is given by $H(X) - H(X/Y)$. The equivocation can be interpreted as the average amount of information, per symbol, lost in transmission. Indeed we see that $H(X/Y) = 0$ if and only if $p(x/y)$ is either 0 or 1, for any x, y, that is, if the reception of a y uniquely specifies the transmitted symbol. When $H(X/Y) = 0$ the channel is called noiseless. If we interpret $H(X)$ as the average amount of information, per symbol, required to specify a given symbol of the ensemble X, with $p(\ )$ as the only initial knowledge about X, then $H(X) - H(X/Y)$ can be considered as the average amount, per symbol transmitted, of the information obtained by the (in general) only partial specification of the transmitted symbol by the received signal.

Let now u(v) represent a sequence of length n (where n is arbitrary but fixed) of statistically independent symbols x(y), and let the space of all sequences be denoted by U(V). In the usual manner we can define the various "product" probabilities. The n will be suppressed throughout. It is now simple to verify the following relations:

$$\log p(u) = \sum_{i=1}^{n} \log p(x_i), \text{ where } u = \{x_1, \ldots, x_n\} \tag{1}$$

$$\log p(u/v) = \sum_{i=1}^{n} \log p(x_i/y_i), \text{ where } v = \{y_1, \ldots, y_n\} \tag{2}$$

$$H(X) = -\frac{1}{n} \sum_U p(u) \log p(u) \tag{3}$$

$$H(X/Y) = -\frac{1}{n} \sum_U \sum_V p(u,v) \log p(u/v) \tag{4}$$

The weak law of large numbers at once gives us the following lemma, which is fundamental for the proof of Shannon's theorem (see also section V).

LEMMA 1. For any ϵ, δ there is an $n(\epsilon, \delta)$ such that for any $n \geq n(\epsilon, \delta)$ the set of u for which the inequality $\left|H(X) + (1/n) \log p(u)\right| < \epsilon$ does not hold has $p(\ )$ probability less then δ. Similarly, but with a different $n(\epsilon, \delta)$, the set of pairs (u, v) for which the inequality $\left|H(X/Y) + (1/n) \log p(u/v)\right| < \epsilon$ does not hold has $p(\ ,\ )$ probability less than δ.

In what follows we shall need only the weaker inequalities $p(u) < 2^{-n(H(X)-\epsilon)}$ and $p(u/v) > 2^{-n(H(X/Y)+\epsilon)}$. The probability of these inequalities failing will be denoted by δ^- and δ^+, respectively.

The following lemma is required to patch up certain difficulties caused by the inequalities of lemma 1 failing to hold everywhere.

LEMMA 2. Let Z be a (u, v) set of $p(\ ,\ )$ probability greater than $1 - \delta_1$ and U_o a set of u with $p(U_o) > 1 - \delta_2$. For each $u \in U$ let A_u be the set of v's such that $(u, A_u) \in Z$. Let $U_1 \subset U_o$ be the set of $u \in U_o$ for which $p(A_u/u) \geq 1 - a$. Then $p(U_1) > 1 - \delta_2 - (\delta_1/a)$.

PROOF. Let U_2 be the set of u for which $p\left(A_u^c/u\right) > a$, where A_u^c is the complement of A_u. Then $p\left(u, A_u^c\right) \geq a p(u)$ for $u \in U_2$, and $\sum_{U_2} \left(u, A_u^c\right)$ is, by the definition of A_u, outside Z. Hence

$$\delta_1 \geq \sum_{U_2} p\left(u, A_u^c\right) \geq a p(U_2), \quad \text{or } p(U_2) \leq \frac{\delta_1}{a}$$

Thus $p(U_2 \cdot U_o) \leq \delta_1/a$ and, using $U_1 = U_o - U_o \cdot U_2$, we have

$$p(U_1) = p(U_o) - p(U_o \cdot U_2) > 1 - \delta_2 - \frac{\delta_1}{a}$$

II. We have seen that, by our definitions, the average amount of information received, per symbol transmitted, is H(X) - H(X/Y). However, in the process of transmission an amount H(X/Y) is lost, on the average. An obvious question is whether it is, in some way, possible to use the channel in such a manner that the average amount of information received, per symbol transmitted, is as near to H(X) - H(X/Y) as we please, while the information lost per symbol is, on the average, as small as we please. Shannon's theorem asserts (1), essentially, that this is possible. More precisely, let there be given a channel with rate H(X) - H(X/Y). Then for any e > 0 and H < H(X) - H(X/Y) there is an n(e,H) such that for each n ≥ n(e,H) there is a family $\{u_i\}$ of message sequences (of length n) of number at least 2^{nH}, and a probability distribution on the $\{u_i\}$ such that, if only the sequences $\{u_i\}$ are transmitted, and with the given probabilities, then they can be detected with average probability of error less than e. The method of detection is that of maximum conditional probability, hence the need for specifying the transmission probability of the $\{u_i\}$. By average probability of error less than e is meant that if e_i is the fraction of the time that when u_i is sent it is misinterpreted, and p_i is u_i's transmission probability, then $\sum_i e_i p_i < e$.

A sufficient condition (2) for the above-mentioned possibility is the following:

For any e > 0 and H < H(X) - H(X/Y) there is an n(e, H) of such value that among all sequences u of length n ≥ n(e, H) there is a set $\{u_i\}$, of number at least 2^{nH}, such that:

1. to each u_i there is a v-set B_i with $p(B_i/u_i) > 1 - e$
2. the B_i are disjoint.

What this says is simply that if we agree to send only the set $\{u_i\}$ and always assume that, when the received sequence lies in B_i, u_i was transmitted, then we shall misidentify the transmitted sequence less than a fraction e of the time. As it stands, however, the above is not quite complete; for, if C is the largest number such that for H < C there is an n(e, H) and a set of at least 2^{nH} sequences u_i satisfying 1 and 2, C is well defined in terms of p(X/Y) alone. However, H(X) - H(X/Y) involves p(X) in addition to p(X/Y). One might guess that C is equal to l.u.b. (H(X) - H(X/Y)) over all choices of p( ). This is indeed so, as the theorem below shows. Note the important fact that we have here a way of defining the channel capacity C without once mentioning information contents or rates. (Strictly speaking we should now consider the channel as being defined simply by p(y/x).) These remarks evidently apply equally well to Shannon's theorem, as we have stated it. We go now to the main theorem.

THEOREM. For any $e > 0$ and $H < C$ there is an $n(e, H)$ such that among all sequences u of length $n \geq n(e, H)$ there is a set $\{u_i\}$, of number at least 2^{nH} such that:

1. to each u_i there is a v-set B_i, with $p(B_i/u_i) > 1 - e$.
2. the B_i are disjoint.

This is not possible for any $H > C$.

PROOF. Let us note here that if we transmit the u_i with equal probability and use a result of section III (namely $P_e \leq e$) we immediately obtain the positive assertion of Shannon's theorem. We shall first indicate only the proof that the theorem cannot hold for $H > C$, which is well known. Indeed if one could take $H > C$ then, as shown in section III one would have, for n sufficiently large, the result that the information rate per symbol would exceed C. But this cannot be (3). Q. E. D. In the following we will take $p(\ )$ as that for which the value C is actually attained (4). We shall see, however, that no use of this fact is actually made in what follows, other than, of course, $C = H(X) - H(X/Y)$.

For given ϵ_1, δ_1^+, ϵ_2, δ_2^-, let $n_1(\epsilon_1, \delta_1^+)$, $n_2(\epsilon_2, \delta_2^-)$ be as in lemma 1 for $p(u/v) > 2^{-n(H(X/Y)+\epsilon_1)}$ and $p(u) < 2^{-n(H(X)-\epsilon_2)}$, respectively. Let us henceforth consider n as fixed and $n \geq \max\left(n_1(\epsilon_1, \delta_1^+), n_2(\epsilon_2, \delta_2^-)\right)$. For Z and U_o in lemma 2 we take, respectively, the sets on which the first two inequalities stated above hold. Then for any $u \in U_1$ (with α as any fixed number $< e$) and v in the corresponding A_u we have:

$$\frac{p(u/v)}{p(u)} > \frac{2^{-n(H(X/Y)+\epsilon_1)}}{2^{-n(H(X)-\epsilon_2)}} = 2^{n(C-\epsilon_1-\epsilon_2)}, \text{ or}$$

$$\frac{p(u, v)}{p(u)} > 2^{n(C-\epsilon_1-\epsilon_2)}\, p(v)$$

Summing v over A_u we have

$$\frac{p(u, A_u)}{p(u)} > 2^{n(C-\epsilon_1-\epsilon_2)}\, p(A_u)$$

Since $1 \geq p(A_u/u)$ we have finally

$$p(A_u) < 2^{-n(C-\epsilon_1-\epsilon_2)}$$

Let $u_1, \ldots, u_N$ be a set M of members of U such that:

a. to each u_i there is a v-set B_i with $p(B_i/u_i) > 1 - e$

b. $p(B_i) < 2^{-n(C-\epsilon_1-\epsilon_2)}$ (See footnote 5.)

c. the B_i are disjoint

d. the set M is maximal, that is, we cannot find a u_{N+1} and a B_{N+1} such that the set $u_1, \ldots, u_{N+1}$ satisfies (a) to (c).

Now for any $u \in U_1$ there is by definition an A_u such that $p(A_u/u) \geq 1 - \alpha > 1 - e$ and as we have seen above, $p(A_u) < 2^{-n(C-\epsilon_1-\epsilon_2)}$. Furthermore, for any $u \in U_1$, $A_u - A_u \cdot \sum_i B_i$ is disjoint from the B_i, and certainly

$$p\left(A_u - A_u \cdot \sum_i B_i\right) < 2^{-n(C-\epsilon_1-\epsilon_2)}$$

If u is not in M, we must therefore have

$$p\left(A_u - A_u \cdot \sum_i B_i/u\right) \leq 1 - e$$

In other words, $p\left(A_u \cdot \sum_i B_i/u\right) \geq e - a$, or certainly

$$p\left(\sum_i B_i/u\right) \geq e - a, \text{ for all } u \in U_1 - M \equiv U_1 - M \cdot U_1$$

Now

$$p\left(\sum_i B_i\right) = \sum_U p\left(\sum_i B_i/u\right) p(u) \geq \left\{\sum_{U_1 - M \cdot U_1} + \sum_{M \cdot U_1}\right\} p\left(\sum_i B_i/u\right) p(u)$$

$$\geq (e-a)\left[1 - \beta_2^- - \frac{\delta_1^+}{a} - p(M \cdot U_1)\right] + (1-e)\, p(M \cdot U_1) \geq (e-a)\left[1 - \delta_2^- - \frac{\delta_1^+}{a}\right]$$

if $e \leq 1/2$, since then $1 - e \geq e - a$.

On the other hand, $p\left(\sum_i B_i\right) < N2^{-n(C-\epsilon_1-\epsilon_2)}$. Hence

$$N2^{-n(C-\epsilon_1-\epsilon_2)} > (e-a)\left[1 - \delta_2^- - \frac{\delta_1^+}{a}\right]$$

If $e > 1/2$ then, using $p(M \cdot U_1) < N2^{-n(H(X)-\epsilon_2)}$, we would obtain

$$N2^{-n(C-\epsilon_1-\epsilon_2)} > (e-a)\left[1 - \delta_2^- - \frac{\delta_1^+}{a} - N2^{-n(H(X)-\epsilon_2)}\right]$$

Since the treatment of both cases is identical, we will consider $e \leq 1/2$.

To complete the proof we must show that for any e and $H < C$ it is possible to choose ϵ_1, ϵ_2, δ_1^+, δ_2^-, $a < e$, and $n \geq \max\left(n_1(\epsilon_1, \delta_1^+),\ n_2(\epsilon_2, \delta_2^-)\right)$ in such a way that the above inequality requires $N \geq 2^{nH}$. Now it is clear that, if, having chosen certain fixed values for the six quantities mentioned, the inequality fails upon the insertion of a given value (say N^*) for N, then the smallest N for which the inequality holds must be greater than N^*. Let us point out that N will in general depend upon the particular maximal set considered.

We take $N^* = 2^{nH}$ and $a = e/2$. Then we can take δ_1^+, δ_2^-, and ϵ_2 so small and n so large that

$$\left[1 - \delta_2^- - \frac{\delta_1^+}{a}\right] \quad \text{is} > \frac{2}{3} \quad \text{say.}$$

We obtain finally $e/3 < 2^{-n(C-H-\epsilon_2-\epsilon_1)}$. Choosing ϵ_2 and ϵ_1 sufficiently small so that $C - H - \epsilon_2 - \epsilon_1 > 0$ we see that for sufficiently large n the inequality $e/3 < 2^{-n(C-H-\epsilon_2-\epsilon_1)}$ fails. Hence for $a = e/2$, for ϵ_1, ϵ_2, δ_1^+, δ_2^- sufficiently small the insertion of $N^* = 2^{nH}$ for N causes the inequality to fail for all n sufficiently large. Thus $N > N^* = 2^{nH}$ for such n. Q.E.D.

It is worthwhile to emphasize that the codes envisaged here, unlike those of Shannon, are uniformly good, i. e., the probability of error for the elements of a maximal set is uniformly <e. These codes are therefore error correcting, which answers in the affirmative the question as to whether the channel capacity can be approached using such codes (6).

If we wish to determine how e decreases as a function of n, for fixed H, we have (7):

$$e \leq a + \frac{A}{B - \left(\delta_1^+/a\right)}, \text{ where } A = 2^{-n(C-H-\epsilon_1-\epsilon_2)}, \quad B = 1 - \delta_2$$

To eliminate the "floating" variable a, we proceed as follows. For $a > 0$

$$a + \frac{A}{B - \left(\delta_1^+/a\right)} \text{ achieves its minimum value at } a_0 = \frac{\left(A\delta_1^+\right)^{1/2} + \delta_1^+}{B}$$

and this value, namely, $\frac{1}{B}\left[A^{1/2} + \left(\delta_1^+\right)^{1/2}\right]^2$, is greater than $\frac{\left(A\delta_1^+\right)^{1/2} + \delta_1^+}{B}$

If we take

$$a = \frac{\left(A\delta_1^+\right)^{1/2} + \delta_1^+}{B} \quad \text{and} \quad e = \frac{1}{B}\left[A^{1/2} + \left(\delta_1^+\right)^{1/2}\right]^2$$

then $a < e$. Hence $\frac{1}{B}\left[A^{1/2} + \left(\delta_1^+\right)^{1/2}\right]^2$ is an upper bound for the minimum value of e which is possible for a given H. This expression is still a function of ϵ_1 and ϵ_2. The best possible upper bound which can be obtained in the present framework is to minimize with respect to ϵ_1 and ϵ_2. This cannot be done generally and in closed form.

Let us remark, however, that at this point we cannot say anything concerning a lower bound for e. In particular, the relation $a < e$ is a condition that is required only if we wish to make use of the framework herein considered.

III. Let us consider a channel (i. e., (S, s), (R, r), p( ) and p(/s) where s is a transmitted and r a received symbol) such that to each s there is an r-set A_s such that $p(A_s/s) \geq 1 - e$ and the A_s are disjoint. For each r let $p_e(r) = 1 - p(s_r/r)$ where s_r is such that $p(s_r/r) \geq p(s/r)$ for all $s \neq s_r$. (Then $p_e(r)$ is simply the probability that when r is received an error will be made in identifying the symbol transmitted, assuming that whenever r is received s_r will be assumed to have been sent.) Now the inequality $\ln a \leq a - 1$ can be used to show that

$$H(S/R) \leq -P_e \log P_e - (1 - P_e) \log (1 - P_e) + P_e \log (N-1)$$

where $P_e = \sum_R p(r)\, p_e(r)$ and N is the number of symbols in S.

We now make use of the special properties of the channel considered. We have

$$P_e = \sum_R p(r)(1 - p(s_r/r)) = 1 - \sum_R p(r)\, p(s_r/r)$$

$$= 1 - \sum_S \sum_{A_s} p(r)\, p(s_r/r) - \sum_{R-\sum_S A_s} p(r)\, p(s_r/r)$$

$$\leq 1 - \sum_S \sum_{A_s} p(r)\, p(s/r) - \sum_{R-\sum_S A_s} p(r)\, p(s_o/r)$$

$$= 1 - \sum_{S-s_o} \sum_{A_s} p(r)\, p(s/r) - \sum_{R-\sum_{S-s_o} A_s} p(r)\, p(s_o/r)$$

$$= 1 - \sum_{S-s_o} p(s)\, p(A_s/s) - p(s_o)\, p\left(R - \sum_{S-s_o} A_s/s_o\right)$$

$$\leq 1 - \sum_{S-s_o} p(s)(1-e) - p(s_o)(1-e) = e \quad \text{where } s_o \text{ is any } s \text{ (8).}$$

Then $H(S/R) \leq -e \log e - (1-e) \log (1-e) + e \log (N-1)$ since for $e < 1/2$ the left side of the above inequality is an increasing function of e. (We assume of course $e < 1/2$.)

Let us consider the elements $u_1, \ldots, u_N$ of some maximal set as the fundamental symbols of a channel. Then regardless of what $p(u_i)$ is, $i = 1, \ldots, N$, the channel is of the type considered above. Hence $P_e \leq e$ (where e is as in II) and

$$H(U/V) \leq -e \log e - (1-e) \log (1-e) + e \log (N-1)$$

Here H(U/V) represents the average amount of information lost per sequence transmitted. The average amount lost per symbol is 1/n H(U/V). Now for $N = 2^{nH}$ and $H < C$, $e = e(n) \to 0$ as $n \to \infty$. Thus $1/n\, H(U/V) \to 0$ as $n \to \infty$. In particular if we take $p(u_i) = 2^{-nH}$, then $1/n\, [H(U) - H(U/V)] \to H$ as $n \to \infty$. (This is the proof mentioned in footnote 2.)

Actually, a much stronger result will be proven, namely, that for $N = 2^{nH}$, $H < C$ (and H fixed, of course) the equivocation per sequence H(U/V), goes to zero as $n \to \infty$. Since $\log (N-1) \approx nH$, a sufficient condition that $H(U/V) \to 0$ as $n \to \infty$ is that $e(n)\, n \to 0$ as $n \to \infty$.

We saw that $e \leq \frac{1}{B}\left[A^{1/2} + \left(\delta_1^+\right)^{1/2}\right]^2$ where $B = 1 - \delta_2$ and $A = 2^{-n(C-H-\epsilon_1-\epsilon_2)}$.

Now if we take ϵ_1, ϵ_2 sufficiently small so that $C - H - \epsilon_1 - \epsilon_2 > 0$ and $H(X) - H - \epsilon_2 > 0$, then the behavior of δ_1^+ as $n \to \infty$ is the only unknown factor in the behavior of e. If the original X consists of only x_1, x_2, and Y consists of only y_1, y_2, and if $p(x_1/y_2) = p(x_2/y_1)$, then $\log p(x/y)$ is only two-valued. If we take $\epsilon_1 = \epsilon(n)$ as vanishing, for $n \to \infty$, faster than $n^{-1/6}$, then a theorem on large deviations (9) is applicable and shows that δ_1^+, and hence e, approaches zero considerably faster than 1/n.

We omit the details inasmuch as a proof of the general case will be given in section V.

IV. Up till now we have considered the set Y of received signals as having a finite number of elements y. One can, however, easily think of real situations where this is not the case, and where the set Y is indeed nondenumerable. Our terminology and notation will follow the supplement of (10).

We define a channel by:

1. the usual set X and a probability distribution p( ) over X
2. a set Ω of points ω
3. a Borel field F of subsets Λ of Ω
4. for each $x \in X$, a probability measure p(/x) on F.

We define the joint probability $p(x, \Lambda) = p(x)\, p(\Lambda/x)$ and $p(\Lambda) = p(X, \Lambda) = \sum_X p(x, \Lambda)$. Since $p(x, \Lambda) \leq p(\ )$ for any x, Λ, we have by the Radon-Nikodym theorem

4.1 $\quad p(x, \Lambda) = \int_\Lambda p(x/\omega)\, p(d\omega)$ where $p(x/\omega)$ may be taken as ≤ 1 for all x, ω.

As the notation implies, $p(x/\omega)$ plays the role of a conditional probability.

We define $H(X) = -\sum_X p(x) \log p(x)$, as before. In analogy with the finite case we define

$$4.2 \quad H(X/Y) = -\sum_X \int_\Omega \log p(x/\omega)\, p(x, d\omega)$$

To show that the integral is finite, we see first, by section 4.1, that

$$p\left(x,\ \{p(x/\omega) = 0\}\right) = 0$$

Furthermore, putting

$$\Lambda_i = \left\{ \frac{1}{2^{i+1}} < p(x/\omega) \leq \frac{1}{2^i} \right\}$$

we have, since $p(\Lambda_i) \leq p(\Omega) = 1$, that

$$\int_{\Lambda_i} p(x/\omega)\, p(d\omega) \leq \frac{p(\Lambda_i)}{2^i} \leq \frac{1}{2^i}$$

Hence

$$4.3 \quad p\left(x, \left\{ \frac{1}{2^{i+1}} < p(x/\omega) \leq \frac{1}{2^i} \right\}\right) \leq \frac{1}{2^i}$$

We therefore have

$$4.4 \quad -\int_\Omega \log p(x/\omega)\, p(x, d\omega) < \sum_{i=0}^{\infty} \frac{i+1}{2^i} < \infty$$

by the ratio test.

Everything we have done in sections I, II, and III can now be carried over without change to the case defined above. A basic theorem in this connection is that we can find a finite number of disjoint sets Λ_j, $\sum_j \Lambda_j = \Omega$ such that $-\sum_X \sum_j p(x, \Lambda_j) \log p(x/\Lambda_j)$ approximates H(X/Y) as closely as desired. Since we make no use of it, we shall not prove it, though it follows easily from the results given above and from standard integral approximation theorems.

V. We shall now show that $e = e(n)$ goes to zero, as $n \to \infty$, faster than $1/n$, which will complete the proof that the equivocation goes to zero as the sequence length $n \to \infty$.

As previously mentioned, it is the behavior of δ_1^+, of lemma 1 that we must determine. The mathematical framework briefly is as follows.

We have the space $X \otimes \Omega$ of all pairs (x, ω) and a probability measure $p(\ ,\ )$ on the measurable sets of $X \otimes \Omega$. We consider the infinite product space $\prod_{i=1}^{\infty} \otimes (X \otimes \Omega)_i$ and the corresponding product measure

$$\prod_{i=1}^{\infty} \otimes p_i(\ ,\ ) \equiv p_\infty(\ ,\ ).$$

Let us denote a "point" of $\prod_{i=1}^{\infty} \otimes (X \otimes \Omega)_i$ by $(x_\infty, \omega_\infty) \equiv \{(x_1, \omega_1), (x_2, \omega_2), \ldots\}$

We define an infinite set of random variables $\{Z_i\}$, $i = 1, \ldots$ on

$$\prod_{i=1}^{\infty} \otimes (X \otimes \Omega)_i$$

by $Z_i(x_\infty, \omega_\infty) = -\log p(x_i/\omega_i)$, that is, Z_i is a function only of the i^{th} coordinate of $(x_\infty, \omega_\infty)$. Clearly the Z_i are independent and identically distributed; we shall put $E(Z_i)$ for their mean value. From section 4.4 we know that the Z_i have moments of the first order. (One can similarly show, using the fact that

$$\infty > \sum_{i=0}^{\infty} \frac{(i+1)^n}{2^i} \quad \text{for any } n > 0,$$

that they have moments of all positive orders.)

Let $S_n = \sum_{i=1}^{n} Z_i$. Then the weak law of large numbers says that for any ϵ_1, δ_1, there is an $n(\epsilon_1, \delta_1)$ such that for $n \geq n(\epsilon_1, \delta_1)$ the set of points $(x_\infty, \omega_\infty)$ on which $\left| \frac{S_n}{n} - E(Z_1) \right| \geq \epsilon_1$ has $p_\infty(\ ,\ )$ measure less than δ_1. Now, in the notation of section I, $S_n(X_\infty, \omega_\infty) = -\log p(u/v)$ where $u = \{x_1, \ldots, x_n\}$ and $v = \{\omega_1, \ldots, \omega_n\}$, while $H(X/Y) = \sum_X \int_\Omega -\log p(x/\omega)\, p(x, d\omega) = E(Z_1)$. What we have stated, then, is simply lemma 1.

Now, we are interested in obtaining an upper bound for

$$\text{Prob}\left\{\frac{S_n}{n} - E(Z_1) \geq \epsilon_1\right\}$$

More precisely we shall find sequences $\epsilon_1(n)$ and $\delta_1^+(n)$ such that, as $n \to \infty$, $\epsilon_1(n) \to 0$, $\delta_1^+(n) \to 0$ faster than $1/n$, and $n(\epsilon_1(n), \delta_1^+(n)) = n$.

Let $Z_i^{(r)} = Z_i$ whenever $Z_i < r$, and $Z_i^{(r)} = 0$ otherwise. By section 4.3, $Z_i^{(r)}$ and Z differ on a set of probability $\leq 1/2^r$. Let $S_n^{(r)} = \sum_{i=1}^{n} Z_i^{(r)}$; then S_n and $S_n^{(r)}$ differ on a set of probability $\leq 1 - (1 - 2^{-r})^n < n/2^r$. Furthermore

$$E(Z_1) - E\left(Z_1^{(r)}\right) \leq \sum_{i=0}^{\infty} \frac{r + 1 + i}{2^{r+i}}$$

by the same argument which led to section 4.4. We thus have:

$$\text{Prob}\left\{\frac{S_n}{n} - E(Z_1) \geq \epsilon_1(n)\right\} \leq \text{Prob}\left\{\frac{S_n^{(r)}}{n} - E(Z_1) \geq \epsilon_1(n)\right\} + \frac{n}{2^r}$$

$$\leq \text{Prob}\left\{\frac{S_n^{(r)}}{n} - E\left(Z_1^{(r)}\right) \geq \epsilon_1(n)\right\} + \frac{n}{2^r},$$

since $E(Z_1) \geq E\left(Z_1^{(r)}\right)$. In order to estimate $\text{Prob}\left\{\frac{S_n^{(r)}}{n} - E\left(Z_1^{(r)}\right) \geq \epsilon_1(n)\right\}$ we use a theorem of Feller (11) which, for our purposes, may be stated as follows:
THEOREM: Let $\{X_i\}$, $i = 1, \ldots, n$ be a set of independent, identically distributed, bounded random variables. Let $S = \sum_{i=1}^{n} X_i$ and let

$$F(x) = \text{Prob}\{S - n\,E(X_1) \leq x\}$$

Put $\sigma^2 = E([X_1 - E(X_1)]^2)$ and take $\lambda > \frac{\sup|X_1 - E(X_1)|}{\sigma n^{1/2}}$. Then if $0 < \lambda x < 1/12$ we have

$$1 - F(x\sigma n^{1/2}) = \exp[-1/2\,x^2\,Q(x)]\,[\{1 - \Phi(x)\} + \theta\lambda \exp(-1/2\,x^2)]$$

where

$$|\theta| < 9, \quad |Q(x)| \leq \frac{1}{7}\left(\frac{12\lambda x}{1 - 12\lambda x}\right) \text{ and } \Phi(x) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{x} \exp[-y^2/2]\,dy$$

In order to apply this theorem, we take $r = r(n)$. Now

$$\sigma\left(Z_1^{(r)}\right) = E\left(\left[Z_1^{(r)} - E\left(Z_1^{(r)}\right)\right]^2\right)^{1/2} \to \sigma(Z_1) \quad \text{as } r \to \infty$$

Hence for suitably large n_o, $\frac{3}{2}\sigma(Z_1) > \sigma\left(Z_1^{(r)}\right) > \frac{1}{2}\sigma(Z_1)$ for $n \geq n_o$. We can now take $\lambda \equiv \lambda(n)$ $\frac{2n^{-1/2}}{\sigma(Z_1)}\,r(n)$.

We henceforth consider $n \geq n_o$. We now have:

$$\text{Prob}\left\{\frac{S_n^{(r)}}{n} - E\left(Z_1^{(r)}\right) \geq \epsilon_1(n)\right\} = \text{Prob}\left\{S_n^{(r)} - n\,E\left(Z_1^{(r)}\right) \geq n\,\epsilon_1(n)\right\}$$

$$= \text{Prob}\left\{S_n^{(r)} - n\,E\left(Z_1^{(r)}\right) \geq \sigma\left(Z_1^{(r)}\right) n^{1/2}\left[n^{1/2}\,\frac{\epsilon_1(n)}{\sigma\left(Z_1^{(r)}\right)}\right]\right\}$$

$$\leq \text{Prob}\left\{S_n^{(r)} - n\,E\left(Z_1^{(r)}\right) \geq \sigma\left(Z_1^{(r)}\right) n^{1/2}\left[n^{1/2}\,\frac{2\epsilon_1(n)}{3(Z_1)}\right]\right\}$$

$$\leq \exp\left[\frac{1}{14}\,x^2\left(\frac{12\lambda x}{1 - 12\lambda x}\right)\right]\left[\{1 - \Phi(x)\} + 9\lambda \exp\left(-\frac{x^2}{2}\right)\right].$$

Using

$$1 - \Phi(x) \sim \frac{1}{(2\pi)^{1/2}\,x} \exp\left(-\frac{x^2}{2}\right)$$

or

$$1 - \Phi(x) \lesssim \frac{2}{(2\pi)^{1/2}\, x} \exp\left(-\frac{x^2}{2}\right)$$

we may rewrite the above as

$$\exp\left(x^2\left[\frac{6\lambda x}{7(1 - 12\lambda x)} - \frac{1}{2}\right]\right) \cdot \left\{9\lambda + \frac{2}{(2\pi)^{1/2}\, x}\right\}$$

Now $\lambda \equiv \lambda(n) = \frac{2n^{1/2}}{\sigma(Z_1)}\, r(n)$ and $x = n^{1/2}\, \frac{2\epsilon_1(n)}{3\sigma(Z_1)}$, while

$$\delta_1^+(n) \leq \exp\left(x^2\left[\frac{6\lambda x}{7(1 - 12\lambda x)} - \frac{1}{2}\right]\right) \cdot \left\{9\lambda(n) + \frac{2}{(2\pi)^{1/2}\, x}\right\} + \frac{n}{2^{r(n)}}$$

It is now clear that we can pick $\epsilon_1(n)$ and $r(n)$ so that $\lambda(n) \to 0$, $x - x(n) \to \infty$, $\lambda x \to 0$ and $\delta_1^+(n) \to 0$ faster than $1/n$.

Let us point out that by using the approximation theorem of section III and thus having to deal with $-\log p(x/\Lambda_j)$, which is bounded, we can eliminate the term $n/2^{r(n)}$. This makes it likely that Feller's theorem can be proven, in our case, without the restriction that the random variables be bounded. There is in fact a remark by Feller that the boundedness condition can be replaced by the condition that Prob $\{|X_i| > n\}$ is a sufficiently rapidly decreasing function of n. But any further discussion would take us too far afield.

VI. We have, up to this point, insisted that the set X of messages be finite. We wish to relax this condition now so that the preceding work can be applied to the continuous channels considered by Shannon (1) and others. However, any attempt to simply replace finite sums by denumerable sums or integrals at once leads to serious difficulties. One can readily find simple examples for which H(X), H(X/Y) and H(X) - H(X/Y) are all infinite.

On the other hand, we may well ask what point there is in trying to work with infinite message ensembles. In any communication system there are always only a finite number of message symbols to be sent, that is, the transmitter intends to send only a finite variety of message symbols. It is quite true that, for example, an atrociously bad telegrapher, despite his intention of sending a dot, dash, or pause, will actually transmit any one of an infinite variety of waveforms only a small number of which resemble intelligible signals. But we can account for this by saying that the "channel" between the telegrapher's mind and hand is "noisy," and, what is more to the point, it is a simple matter to determine all the statistical properties that are relevant to the capacity of this "channel." The channel whose message ensemble consists of the finite number of "intentions" of the telegrapher and whose received signal ensemble is an infinite set of waveforms resulting from the telegrapher's incompetence and noise in the wire is thus of the type considered in section IV.

The case in which one is led to the consideration of so-called continuous channels is typified by the following example. In transmitting printed English via some teletype system one could represent each letter by a waveform, or each pair by a waveform, or every letter and certain pairs by a waveform, and so on. We have here an arbitrariness both in the number of message symbols and in the waveforms by which they are to be represented. It is now clear that

we should extend the definition of a channel and its capacity in order to include the case given above.

DEFINITION. Let X be a set of points x and Ω a set of points ω. Let F be a Borel field of subsets Λ of Ω, and let $p(\ /x)$ be, for each $x \in X$, a probability measure on F. For each finite subset R of X the corresponding channel and its capacity C_R is well defined by section IV. The quantity $C = \text{l.u.b. } C_R$ over all finite subsets R of X will be called the capacity of the channel $\{X, p(\ /x), \Omega\}$.

Now for any $H < C$ there is a C_R with $H < C_R \leq C$, so that all our previous results are immediately applicable.

We shall now show that the channel capacity defined above is, under suitable restrictions, identical with that defined by Shannon (1).

Let X be the whole real line, and Ω, ω, F, and Λ as usual. Let $p(x)$ be a continuous probability density over X and for each $\Lambda \in F$, let $p(\Lambda/x)$ satisfy a suitable continuity condition. (See the Appendix for this and subsequent mathematical details.) Then $p(\Lambda) \equiv \int_{-\infty}^{\infty} p(x)p(\Lambda/x)\,dx$ is a probability measure. Since $p(x, \Lambda) \equiv p(x)p(\Lambda/x)$ is, for each x, absolutely continuous with respect to $p(\Lambda)$ we can define the Radon-Nikodym derivative $p(x/\omega)$ by $p(x, \Lambda) = \int_{\Lambda} p(x/\omega)p(d\omega)$. Then, with the x-integral taken as improper, we can define

$$C_p \equiv \int_{-\infty}^{\infty} dx \int_{\Omega} p(x, d\omega) \log \frac{p(x/\omega)}{p(x)} \geq 0$$

If we put $C_s = \text{l.u.b. } C_p$ over all continuous probability densities $p(x)$, then C_s is Shannon's definition of the channel capacity. The demonstration of the equivalence of C, as defined above, and C_s is now essentially a matter of approximating an integral by a finite sum, as follows:

If C_s is finite, then we can find a C_p arbitrarily close to C_s; if $C_s = +\infty$ we can find C_p arbitrarily large. We can further require that $p(x)$ shall vanish outside a suitably large interval, say $[-A, A]$. We can now find step-functions $g(x)$ defined over $[-A, A]$ that approximate $p(x)$ uniformly to any desired degree of accuracy, and whose integral is 1. For such a step-function, C_g is well defined and approximates C_p as closely as desired by suitable choice of $g(x)$.

Let $g(x)$ have n steps, with area p_i, and of course $\sum_{1}^{n} p_i = 1$. By suitably choosing positive numbers a_{ij}, integers N_i, and points x_{ij}, with x_{ij} lying in the i^{th} step of $g(x)$ and $\sum_{j=1}^{N_i} a_{ij} = p_i$, we can approximate

$$p(\Lambda) \equiv \int_{-A}^{A} g(x)\, p(\Lambda/x)\, dx \quad \text{by} \quad \sum_{i=1}^{n} \sum_{j=1}^{N_i} a_{ij}\, p(\Lambda/x_{ij})$$

and hence C_g by C_R, where $R = \{x_{ij}\}$. Thus $C \geq C_s$. On the other hand, let $R = \{x_i\}$, not as taken above. Let $p(x_i)$ be such that $H(X) - H(X/Y) = C_R$. Then the singular function $\sum_i p(x_i)\, \delta(x - x_i)$, where $\delta(\ )$ is the Dirac delta-function, can be approximated by continuous probability densities $p(x)$ such that C_p approximates C_R. Hence $C_s \geq C$, or $C = C_s$.

This can clearly be generalized to the case in which X is n-dimensional Euclidean space.

VII. We now wish to relax the condition of independence between successive transmitted symbols. Our definitions will be those of Shannon, as generalized by McMillan, whose paper (1) we now follow.

By an alphabet we mean a finite abstract set. Let A be an alphabet and I the set of all integers, positive, zero, and negative. Denote by A^I the set of all sequences $x = (\ldots, x_{-1}, x_0, x_1, \ldots)$ with $x_t \in A$, $t \in I$.

A cylinder set in A^I is a subset of A^I defined by specifying an integer $n \geq 1$, a finite sequence $a_0, \ldots, a_{n-1}$, of letters of A, and an integer t. The cylinder set corresponding to these specifications is $\left\{x \in A^I / x_{t+k} = a_k,\ k = 0, \ldots, n-1\right\}$. We denote by F_A the Borel field generated by the cylinder sets.

An information source $[A, \mu]$ consists of an alphabet A and a probability measure μ defined on F_A. Let T be defined by $T(\ldots, x_{-1}, x_0, x_1, x_2, \ldots) = (\ldots, x'_{-1}, x'_0, x'_1 \ldots)$ where $x'_t = x_{t+1}$. Then $[A, \mu]$ will be called stationary if, for $S \in F_A$, $\mu(S) = \mu(TS)$ (clearly T preserves measurability) and will be called ergodic if it is stationary and S = TS implies that $\mu(S) = 1$ or 0.

By a channel we mean the system consisting of:

1. a finite alphabet A and an abstract space B.
2. a Borel field of subsets of B, designated by β, with $B \in \beta$
3. the Borel field of subsets of $B^I \equiv \prod_{-\infty}^{\infty} \otimes B_i$ (where $B_i = B$) which we define in the usual way, $\prod_{-\infty}^{\infty} \otimes \beta$, and designate F_β.
4. a function ν_x which is, for each $x \in A^I$, a probability measure on F_β, and which has the property that if $x_t^1 = x_t^2$ for $t \leq n$, then $\nu_{x^1}(S) = \nu_{x^2}(S)$ for any $S \in F_\beta$ of the form $S = S_1 \otimes S_2$, where $S_1 \in \prod_{-\infty}^{n} \otimes \beta$ and $S_2 = \prod_{n+1}^{\infty} \otimes B$.

Consider a stationary channel whose input A is a stationary source $[A, \mu]$. Let $C^I = A^I \otimes B^I$ and $F_C = F_A \otimes F_\beta$. We can define a probability measure on F_C by $p(R, S) \equiv p(R \otimes S) = \int_R \nu_x(S)\, d\mu(x)$ for $R \in F_A$, $S \in F_\beta$, assuming certain measurability conditions for $\nu_x(S)$. It is then possible to define the information rate of the channel source, the equivocation of the channel, and the channel capacity in a manner analogous to that of section I. Assuming that $\mu(\ )$ and $p(\ ,\ )$ are ergodic, McMillan proves lemma 1 of section I in this more general framework. Hence the proof of section III remains completely valid, except for the demonstration that the theorem cannot hold for H > C.

The difficulty that we wish to discuss arises in the interpretation of $p(\ /u)$. A glance at McMillan's definitions shows that p(B/u) no longer can be interpreted as "the probability of receiving a sequence lying in B, given that the sequence u was sent." This direct causal interpretation is valid only for $\nu_x(\ )$. But the result of the theorem of section II is the existance of a set u_i and disjoint sets B_i such that $p(B_i/u_i) > 1 - e$. Under what conditions can we derive from this an analogous statement for $\nu_{u_i}(B_i)$?

Suppose that for a given integer N we are given, for each sequence $x_1, \ldots, x_{N+1}$ of message symbols, a probability measure $\nu(\ /x_1, \ldots, x_{N+1})$

on the Borel field β of received signals (not sequences of signals). We envisage here the situation in which the received signal depends not only upon the transmitted symbol x_{N+1} but also upon the preceding N symbols which were transmitted.

If $u = \{x_1, \ldots, x_n\}$ then

$$p(\ /u) \equiv \sum_{[x_{-N+1}, \ldots, x_o]} \frac{p(x_{-N+1}, \ldots, x_n)}{p(x_1, \ldots, x_n)} \times [\nu(\ /x_{n-N}, \ldots, x_n) \otimes \ldots \otimes \nu(\ /x_{-N+1}, \ldots, x_1)]$$

Let us write the bracket term, which is a probability measure on received sequences of length n, as $\nu_n(\ /x_{-N+1}, \ldots, x_n)$. Now if $p(B_i/u_i) > 1 - e$, then, since

$$\sum_{[x_{-N+1}, \ldots, x_o]} \frac{p(x_{-N+1}, \ldots, x_n)}{p(x_1, \ldots, x_n)} = 1$$

there must be at least one sequence $\{x_{-N+1}, \ldots, x_n\}$ for which

$$\nu_n(B_i/x_{-N+1}, \ldots, x_n) > 1 - e$$

A minor point still remains: we had 2^{nH} sequences u_i and we now have the same number of sequences, but of length n + N. In other words, we are transmitting at a rate H' = (n/n+N) H. But since N is fixed we can make H' as near as we choose to H by taking n sufficiently large; hence we can still transmit at a rate as close as desired to the channel capacity.

It is evident that by imposing suitable restrictions on $\nu_x(\ )$ we can do the same sort of thing in a more general context. These restrictions would amount to saying that the channel characteristics are sufficiently insensitive to the remote past history of the channel.

In this connection some interesting mathematical questions arise. If we define the capacity following McMillan for the $\nu(\ /x_1, \ldots, x_{N+1})$ as above, is the capacity actually achieved? It seems reasonable that it is, and that the channel source that attains the capacity will automatically be of the mixing type (see ref. 12, p. 36, Def. 11.1; also p. 57) and hence ergodic. Because of the special form of $\nu_x(\ )$ it easily follows that the joint probability measure would likewise be of mixing type and hence ergodic.

The question of whether or not the equivocation vanishes in this more general setup is also unsettled. Presumably one might be able to extend Feller's theorem to the case of nonindependent random variables that approach independence, or perhaps actually attain independence when far enough apart. To my knowledge nothing of this sort appears in the literature.

Finally there is the question of whether or not, in the more general cases, the assertion that for H > C the main theorem cannot hold is still true. While this seems likely, at least in the case of a channel with finite memory, it is to my knowledge unproven.

APPENDIX

It is our purpose here to supply various proofs that were omitted in the body of the work.

1. $H(X) - H(X/Y)$ is a continuous function of the $p(x_i)$, $i = 1, \ldots, a$.

PROOF. $H(X)$ is clearly continuous. To show the same for $H(X/Y)$ we need only show that for each i, $-p(x_i) \int_\Omega \log p(x_i/\omega)\, p(d\omega/x_i)$ is a continuous function of $p(x_1), \ldots, p(x_a)$. Now

$$p(x_i/\omega) = \frac{p(x_i, d\omega)}{p(d\omega)} = p(x_i) \frac{p(d\omega/x_i)}{p(d\omega)}$$

But since $\sum_i p(\Lambda/x_i) \geq p(\Lambda)$, we have (see ref. 13, p. 133)

$$\frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} = \frac{p(d\omega/x_i)}{p(d\omega)} \cdot \frac{p(d\omega)}{\sum_i p(d\omega/x_i)}$$

$$= \frac{p(d\omega/x_i)}{p(d\omega)} \cdot \frac{\sum_i p(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)}$$

almost everywhere with respect to $\sum_i p(\ /x_i)$ and hence, certainly, almost everywhere with respect to each $p(\ /x_i)$. Thus

$$\frac{p(d\omega/x_i)}{p(d\omega)} = \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \Bigg/ \frac{\sum_i p(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)}$$

almost everywhere with respect to $p(\ )$. The dependence on the $p(x_i)$ is now explicitly continuous, so that each $p(x_i/\omega)$ is a continuous function of $p(x_1), \ldots, p(x_a)$ almost everywhere with respect to each $p(\ /x_i)$. We now wish to show that $-p(x_i) \int_\Omega \log p(x_i/\omega)\, p/d\omega(x_i)$ is a continuous function of the $p(x_i)$.

To this end let $\{p_j(x_1), \ldots, p_j(x_a)\}$, $j = 1, \ldots$ be a convergent sequence of points in a-dimensional Euclidean space R_a, with limit $\{p_o(x_1), \ldots, p_o(x_a)\}$. Then we have $\lim_{j\to\infty} p_j(x_i/\omega) = p_o(x_i/\omega)$ almost everywhere with respect to each $p(\ /x)$. We must now show that

$$-p_j(x_i) \int_\Omega \log p_j(x_i/\omega)\, p(d\omega/x_i) \to -p_o(x_i) \int_\Omega \log p_o(x_i/\omega)\, p(d\omega/x_i).$$

Suppose, first, that $p_o(x_i) \neq 0$. Now from section IV we have

$$\int_\Omega -\log \left\{ \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \Bigg/ \frac{\sum_i p(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \right\} p(d\omega/x_i) < \infty$$

whenever $p(x_i) \neq 0$. Take $p(x_i) = p(x_2) = \ldots p(x_a) = 1/a$. Then

$$\int_\Omega -\log a \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)}\, p(d\omega/x_i) < \infty \quad \text{or clearly}$$

$$\int_\Omega -\log \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)}\, p(d\omega/x_i) < \infty. \quad \text{But}$$

$$-\log \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \geq -\log \left\{ \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \Big/ \frac{\sum_i p_j(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \right\}, \quad j = 1, 2, \ldots$$

Since the last term is also bounded below by $\log p_j(x_i)$, then by reference 14, p. 110, we have

$$\lim_{j\to\infty} \int_\Omega -\log \left\{ \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \Big/ \frac{\sum_i p_j(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \right\} p(d\omega/x_i)$$

$$= \int_\Omega -\log \left\{ \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \Big/ \frac{\sum_i p_j(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \right\} p(d\omega/x_i)$$

Since $p_o(x_i) \neq 0$, $-p_j(x_i) \int_\Omega \log p(x_i/\omega)\, p(d\omega/x_i) = p_j(x_i)$

$$\int_\Omega -\log \left[p_j(x_i) \cdot \left\{ \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \Big/ \frac{\sum_i p_j(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \right\} \right] p(d\omega/x_i) \to p_o(x_i)$$

$$\int_\Omega -\log \left[p_o(x_i) \left\{ \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \Big/ \frac{\sum_i p_o(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \right\} \right] p(d\omega/x_i)$$

$$= -p_o(x_i) \int_\Omega \log p(x_i/\omega)\, p(d\omega/x_i)$$

If $p_o(x_i) = 0$, we can clearly assume $p_j(x_i) \neq 0$, since we have to show that $-p_j(x_i) \int_\Omega \log p_j(x_i/\omega)\, p(d\omega/x_i) \to 0$. As before we have

$$-\log \frac{p_j(x_i/\omega)}{p_j(x_i)} \leq -\log \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)}, \text{ therefore}$$

$$-p_j(x_i) \int_\Omega \log p_j(x_i/\omega)\, p(d\omega/x_i) \leq p_j(x_i) \int_\Omega -\log \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)}\, p(d\omega/x_i)$$

$$+ p_j(x_i) \log \frac{1}{p_j(x_i)} \to 0 \quad \text{as } p_j(x_i) \to 0 \text{ (i.e., as } j \to \infty).$$

2. We wish here to rigorize the discussion of section VI.

We assume that $p(\ /x)$ satisfies the following continuity condition: For any finite closed interval I and any ϵ there is a $\delta(I, \epsilon)$ such that

$$\left| \frac{p(\Lambda/x_2)}{p(\Lambda/x_2)} - 1 \right| < \epsilon \quad \text{for} \quad |x_1 - x_2| \leq \delta \quad \text{and } x_1, x_2 \in I,$$

whenever $p(\Lambda/x_2) \neq 0$. It follows that if, for $x_1 \in I$, $p(\Lambda/x_1) = 0$, then for $x_2 \in I$ and $|x_1 - x_2| < \delta$, $p(\Lambda/x_2) = 0$. (Indeed, since $\{x/p(\Lambda/x) = 0\}$ is evidently both open and closed, for any Λ, $p(\Lambda/x)$ either vanishes everywhere or nowhere.) That $p(\Lambda) \equiv \int_{-\infty}^{\infty} p(x)\, p(\Lambda/x)\, dx$ is a probability measure is a simple consequence of reference 14, p. 112, Theorem B. Since $p(x)\, p(\Lambda/x)$ is continuous, $p(\Lambda)$ can vanish only if $p(x)\, p(\Lambda/x)$ is zero for all x. Hence,

for all x, $p(x)\,p(\Lambda/x)$ is absolutely continuous with respect to $p(\Lambda)$.

We can sharpen this result as follows: Let I be a closed interval over which $p(x) \neq 0$. Then for a given ϵ we can find a δ such that $p(x_1) > p(x_2)/2$ and $p(\Lambda/x_1) \geq (1-\epsilon)\,p(\Lambda/x_2)$, for $x_1, x_2 \in I$ and $|x_1 - x_2| < \delta$. We thus have

$$\int_{-\infty}^{\infty} p(x)\,p(\Lambda/x)\,dx \geq 2\delta\,\frac{p(x_2)}{2}\,p(\Lambda/x_2)(1-\epsilon) = \delta\,p(x_2)(1-\epsilon)\,p(\Lambda/x_2)$$

Thus for any $x_2 \in I$,

$$p(x_2)\cdot p(\Lambda/x_2) \leq \frac{1}{(1-\epsilon)\delta}\,p(\Lambda) \equiv k(x_2)\,p(\Lambda)$$

which defines $k(x_2) < \infty$. As in section IV, we can easily show that

$$-\infty < -\int_{\Omega} \log p(x/\omega)\,p(x, d\omega) < \infty \quad \text{for all } x. \quad \text{Now}$$

$$\int_{\Omega} p(x, d\omega) \log \frac{p(x)}{p(x/\omega)} \leq \int_{\Omega} p(x, d\omega) \left\{ \frac{p(x)}{p(x/\omega)} - 1 \right\} \log e$$

$$= \int_{\Omega} p(x/\omega)\,p(d\omega) \left\{ \frac{p(x)}{p(x/\omega)} - 1 \right\} \log e = 0,$$

the next to last equality being justified by reference 14, p. 133. Therefore, if $\int_{\Omega} p(x, d\omega) \log \dfrac{p(x/\omega)}{p(x)}$ is, say, continuous in x, then

$$\lim_{a \to \infty} \int_{-a}^{a} \int_{\Omega} p(x, d\omega) \log \frac{p(x/\omega)}{p(x)}\,dx$$

is meaningful and is either positive or equal to $+\infty$.

We shall now show that $\int_{\Omega} p(x, d\omega) \log \dfrac{p(x/\omega)}{p(x)}$ is indeed continuous. To this end let x_i be a convergent sequence of real numbers with limit x_o. We shall show that $p(x_i/\omega) \to p(x_o/\omega)$ almost everywhere with respect to $p(x_o, \ )$. (Since for $p(x_o) = 0$ this assertion is trivially true, we assume that $p(x_o \neq 0)$.) Let $A_{in}^{+} = \{p(x_i/\omega) - p(x_o/\omega) > 1/n\}$ and $A_{in}^{-} = \{p(x_i/\omega) - p(x_o/\omega) < -1/n\}$. Now

$$p(x_i)\,p(A_{in}^{+}/x_i) - p(x_o)\,p(A_{in}^{+}/x_o) = \int_{A_{in}^{+}} (p(x_i/\omega) - p(x_o/\omega))\,p(d\omega) > 1/n\,p(A_{in}^{+}).$$

There is clearly no loss in generality in assuming $p(x_i) \neq 0$. Then

$$p\left(A_{in}^{+}/x_i\right) - p\left(A_{in}^{+}/x_o\right) > \frac{p(x_o)}{k(x_o)\,n\,p(x_i)}\,p\left(A_{in}^{+}/x_o\right) + \frac{p(x_o) - p(x_i)}{p(x_i)}\,p\left(A_{in}^{+}/x_o\right).$$

Now $\dfrac{p(k_o)}{k(x_o)\,n\,p(x_i)} + \dfrac{p(x_o) - p(x_i)}{p(x_i)}$ is positive and bounded away from zero for all i sufficiently large. By the continuity condition on $p(\ /x)$ we therefore have $p\left(A_{in}^{+}/x_o\right) = 0$ for $i > i(n)$ suitably chosen. We get a similar result for $p\left(A_{in}^{-}/x_o\right)$. Let A_n^{+} be the set of points ω which lie in infinitely many A_{in}^{+}, and similarly for A_n^{-}. Then $p\left(A_n^{\pm}/x_o\right) = 0$, and so,

$$p\left(\sum_n A_n^+ + \sum_n A_n^-/x_o\right) = p\left(x_o, \sum_n \left(A_n^+ + A_n^-\right)\right) = 0$$

But for any $\omega \in \Omega - \sum_n A_n^+ - \sum_n A_n^-$, $p(x_i/\omega) \to p(x_o/\omega)$, which was to be shown.

As before, let x_i be a convergent sequence with limit x_o.

a. Let us assume first that $p(x_o) \neq 0$. Now

$$\left|\int_\Omega -\log p(x_o/\omega)\, p(d\omega/x_o) - \int_\Omega -\log p(x_i/\omega)\, p(d\omega/x_i)\right|$$

$$= \left|\int_\Omega \left[-\log p(x_o/\omega) + \log p(x_i/\omega)\right] p(d\omega/x_o) - \int_\Omega -\log p(x_i/\omega)\, p(d\omega/x_i)\right.$$

$$\left. + \int_\Omega -\log p(x_i/\omega)\, p(d\omega/x_o)\right| \leq \left|\int_\Omega \left[-\log p(x_o/\omega) + \log p(x_i/\omega)\right] p(d\omega/x_o)\right|$$

$$+ \left|\int_\Omega -\log p(x_i/\omega)\, p(d\omega/x_o) - \int_\Omega -\log p(x_i/\omega)\, p(d\omega/x_i)\right|$$

To show that the first of the last two terms goes to zero, we remark, first, that since $p(x_o) \neq 0$ and $p(\Lambda/x_o) \leq (1+\alpha)\, p(\Lambda/x_i)$ for any Λ, for i suitably large, it follows, as in section IV, that

$$\int_{\{-\log p(x_i/\omega) \geq R\}} -\log p(x_i/\omega)\, p(d\omega/x_o)$$

is uniformly bounded for all i, where we use the previously shown result that $p(x)\, p(\Lambda/x) \leq k(x)\, p(\Lambda) \leq M\, p(\Lambda)$ for M suitably chosen and x in a closed interval containing x_o. It is now a simple exercise, by using reference 14, p. 110, to justify the interchange of limit and integration, so that the term in question vanishes as $i \to \infty$. The relation $p(\ /x_o) \leq (1+\alpha)\, p(\ /x_i) \leq (1+\alpha)^2\, p(\ /x_o)$, with $\alpha \to 0$ as $i \to \infty$, at once shows that the second term likewise vanishes as $i \to \infty$.

b. Now suppose that $p(x_o) = 0$. Then by definition we take

$$\int_\Omega -\log p(x_o/\omega)\, p(x_o, d\omega) = 0$$

If $p(x)$ is identically zero in some neighborhood of x_o there is nothing to be proven. We can then assume that $p(x_i) \neq 0$. For a closed interval containing x_o and the x_i, we have, for $|x_i - x_j|$ sufficiently small (or equivalently, i and j sufficiently large) that $p(\ /x_i) \geq (1-\epsilon)\, p(\ /x_j)$. Thus

$$p(x_i/\omega) \geq p(x_i)\, \frac{p(d\omega/x_j)}{p(d\omega)}\, (1-\epsilon). \quad \text{Hence}$$

$$-\log p(x_i/\omega) \leq -\log\left[p(x_i)(1-\epsilon)\right] - \log \frac{p(d\omega/x_j)}{p(d\omega)}$$

for fixed j and any i, both sufficiently large. Further, since $p(x_j) \neq 0$, $p(x_j)\, p(\Lambda/x_j) = M\, p(\Lambda)$ for suitable M. Hence

$$p(x_i)\, p(\Lambda/x_i) \leq \frac{1}{1-\epsilon} \frac{p(x_i)}{p(x_i)}\, M\, p(\Lambda)$$

Since $p(x_i) \to 0$, we have, for sufficiently large i, $p(x_i)\, p(\Lambda/x_i) \leq p(\Lambda)$, so that $p(x_i/\omega) \leq 1$ or $-\log p(x_i/\omega) \geq 0$. Therefore

$$\int_\Omega - \log p(x_i/\omega)\, p(x_i, d\omega) \leq - p(x_i) \log [p(x_i)(1-\epsilon)]$$

$$+ p(x_i) \int_\Omega - \log \frac{p(d\omega/x_j)}{p(d\omega)}\, p(d\omega/x_i)$$

As i approaches ∞, the last integral approaches

$$\int_\Omega - \log \frac{p(d\omega/x_j)}{p(d\omega)}\, p(d\omega/x_o), \quad \text{which is} < \infty,$$

using arguments as in section IV. Since $p(x_i) \to 0$, we have, finally,

$$\int_\Omega - \log p(x_i/\omega)\, p(x_i, d\omega) \to 0 \quad \text{as } i \to \infty.$$

References and Footnotes

1. C. E. Shannon, A mathematical theory of communication, Bell System Tech. J. 27, 379-423, 623-656; also B. McMillan, The basic theorems of information theory, Ann. Math. Stat. 24, 196-219.

2. That it is indeed sufficient will be shown in section III.

3. R. M. Fano, Lecture notes on statistical theory of information, Massachusetts Institute of Technology, spring, 1952. This statement asserts that if the channel is considered as transmitting sequence by sequence its capacity per symbol is still bounded by C. Using the fact that the reception rate per symbol may be written as

$$\frac{H(V) - H(V/U)}{n}$$

the statement follows upon noticing that H(V/U) depends only upon single-received-symbol probabilities and that H(V) is a maximum when those probabilities are independent. The expression H(V) - H(V/U) then reduces to a sum of single-symbol channel rates, from which the assertion follows at once.

4. It is not difficult to see that H(X) - H(X/Y) is a continuous function of the "variables" $r_i = p(x_i)$, $i = 1, \ldots, \alpha$. This is true also in the context of section IV (c.f. Appendix). Since the set of points in α-dimensional cartesian space R_α defined by $r_i \geq 0$ and $\sum_{i=1}^{\alpha} r_i = 1$ is a closed set, H(X) - H(X/Y) attains a maximum value. This point is, however, not critical, for, given $H < C$ we can certainly find p( ) such that $H < H(X) - H(X/Y) < C$ and then use H(X) - H(X/Y) in place of C.

5. This condition appears to be superfluous. It is, however, strongly indicated by the immediately preceding result and is, in fact, essential for the proof.

6. E. M. Gilbert, A comparison of signalling alphabets, Bell System Tech. J. 31, in particular p. 506.

7. Up to here, the possibility that certain quantities are not integers can be seen not to invalidate any of the various inequalities. In what follows, the modifications needed to account for this possibility are obvious and insignificant and are therefore omitted.

8. Word-wise, this string of inequalities states simply: (a) that in order to minimize the probability of misidentifying the transmitted s we should guess the s with greatest conditional probability as the one actually transmitted; (b) if instead of the above recipe, we assume that s was sent, whenever $r \epsilon A_s$ is received, for all s except s_o, and that in all other circumstances we shall assume s_o to have been sent, then the probability of error is less than e; (c) hence, since P_e is the error obtained by the best method of guessing, $P_e \leq e$.

9. See reference 13, pp. 144-5. This was pointed out by Professor R. M. Fano.

10. J. L. Doob, Stochastic Processes (John Wiley and Sons, Inc., New York, 1953).

11. W. Feller, Generalization of a probability limit theorem of Cramer, Trans. Amer. Math. Soc. 54, 361-372 (1943).

12. E. Hopf, Ergodentheorie (Julius Springer, Berlin, 1937).

13. W. Feller, An Introduction to Probability Theory (John Wiley and Sons, Inc., New York, 1950).

14. P. R. Halmos, Measure Theory (D. Van Nostrand, New York, 1950).

BINARY CODING

Marcel J. E. Golay
Signal Corps Engineering Laboratories
Fort Monmouth, New Jersey

INTRODUCTION

The upper bound given by Shannon[1] to the transmission capacity of a noisy, discrete channel has challenged the mathematicians, who have accepted this challenge, to devise digital error correcting codes or coding systems approximating as close as possible this upper bound.

This mathematical effort has been concentrated in the binary system and has had the aim to devise codes which are as efficient as possible, in the sense that, given an upper limit to the number m of errors during the transmission and reception of a block or message of n binits, the following obtains: (a) All messages are received in all cases without equivocation; (b) The number of transmittable messages approaches as close as possible the value $2^n / \sum_{m=0}^{m=e} \frac{n!}{(n-m)!\,m!}$, the sum in the denominator being the sum of the (e + 1)st numbers of the nth line of Pascal's triangle, and representing the number of ways in which any one transmitted message can be received when transmission errors in any number from zero to e can occur.

Codes in which the upper limit is exactly reached will be termed lossless codes in what follows, and it may be worth noting here the paradoxical circumstance that while the existence proof for codes approaching indefinitely Shannon's upper bound was based on the assumption of codes consisting of random messages, the search for efficient or lossless codes has been successful to the extent that codes were discovered which were characterized by deeply seated, entwined symetries.

It is the purpose of this discussion to explore certain aspects of this circumstance, and to describe some group-theoretical approaches to coding problems.

The first example of a symbol correcting code was given by Shannon[2] who quotes Hamming's lossless coding of a seven binit message, none or one of which can be received in error. This case was extended by the writer to blocks of 2^n-1 binary symbols, and, more generally, to blocks of $\frac{p^n-1}{p-1}$ p-nary symbols (p prime), none or one of which can be received in error.[3] With the exception of the trivial cases of (2n + 1) binit messages, up to n of which can be received in error, and of two special cases treated in the last paper cited, these are the only cases of lossless symbol coding known, and the possibility must be considered that others do not exist. Their impossibility will be demonstrated below for the case of lossless 2-error correcting symbol codes, and it will be shown also that the search for e-error correcting symbol codes need be a finite one only, because lossless symbol correcting codes become impossible beyond a determinable message length, for any one selected value of e.

These results will leave open the question of whether cases of two or more error-correcting lossless message codes exist (outside of the one mentioned above) because message codes form a more general class of codes than symbol codes, which form a sub-class of it only, and various examples of message codes which are more efficient than symbol codes will be cited, and their mode of formation illustrated. It is this mode of formation which is suggestive of the kind of group theoretical approach which the writer believes to be the most promising for the class of coding problems considered.

Symbol Correcting Codes

When a symbol correcting digital code exists for the transmission of n-binit messages, up to e of which can be received in error, and i of which (the X_m's) carry the message while the remaining j = n - i (the Y_k's) binits are redundant and are provided to remove the equivocation, the transmitted binits are related by the matrix:

$$E_m \equiv \sum_{k=1}^{k=j} a_{mk} Y_k + X_m \equiv 0 \pmod 2, \quad m = 1, 2, \ldots i$$

and the essential property of this matrix is that the E's recalculated from the partially erroneously received X_k's and Y_m's form a j-binit number E, which will be termed the corrector, and which determines univocally which symbols were received in error.

1 Bell System Technical Journal, July, 1948.

2 Loc. cit., p. 418.

3 Marcel J. E. Golay, "Notes on Digital Coding," Proc. I.R.E., vol. 37, p. 637; 1949.

When the code is lossless, a first condition must obtain, which stipulates that all possible cases of up to e errors are represented by all the possible values of the corrector:

$$\sum_{k=0}^{k=e} \frac{n!}{(n-k)!\,k!} = 2^j \qquad (1)$$

Another condition can be obtained as follows:

Whenever all but one Y_k and all X_m's received are zero, the binits of the corrector E(k) consist of the series of a_{mk} values for the particular k considered, and will be termed the characteristic of Y_k.

Whenever all but one X_m and all Y_k's are zero, the binits of the corrector E(m) consist of zeroes with a single one corresponding to the particular m considered, and will be called the characteristic of X_m. In general, the corrector E consists of the j-binit number formed by adding modulo 2 the corresponding binits of the characteristics of the symbols received in error. A general condition for a lossless code is that all possible (boolean) additions thus made of up to e characteristics reproduce exactly, and only once, each of the 2^j possible values of the corrector.

If the parity of the characteristics or of the corrector is defined as zero when the number of ones in these numbers is even, and as one otherwise, it will be readily seen that the parity of the corrector will be the parity of the number of odd characteristics (parity one) required to form it. In a lossless code, all even correctors, 2^{j-1} in number, shall be formed from all possible additions of up to e characteristics in which the number of odd characteristics employed, 2s, is always even. Let r be the total number of odd characteristics. We shall have the other condition sought:

$$\sum_{k=0}^{k=e} \sum^{s} \frac{(n-r)!}{(n-r-k+2s)!(k-2s)!} \cdot \frac{r!}{(r-2s)!(2s)!} = 2^{j-1} \qquad (2)$$

A corollary from (1) and (2) can be obtained by subtracting the second relation from the first, member by member. This operation yields the relation:

$$\sum_{k=0}^{k=e} \sum^{s} \frac{(n-r)!}{(n-k-r+2s-1)!(k-2s+1)!} \cdot \frac{r!}{(r-2s+1)!(2s-1)!} = 2^{j-1} \qquad (3)$$

When e = 2, (1) and (2) can be written:

$$n^2 + n - 2 = 2^{j+1} \qquad (1a)$$

$$(n - r + 1)\ r = 2^{j-1} \qquad (2a)$$

These relations are satisfied for n = 5 and r = 2 or 4 (r = 2 does not correspond to any code, and r = 4 corresponds to the trivial case of a five binit message, up to two of which can be in error), but for larger values of n and r the approximation obtained by eliminating all but the highest degree terms in the left members of (1a) and (2a).

$$n^2 = 2^{j+1} \qquad (1b)$$

$$(n - r)\ r = 2^{j-1} \qquad (2b)$$

indicates that $n \simeq 2r$.
(2a) requires that:

$$r = 2^{\frac{j-1}{2}} \quad \text{and} \quad n + 1 = 2^{\frac{j+1}{2}}$$

and substitution of the value for n derived from the last relation in (1a) yields n = 1, which contradicts the postulation of a large n.

In the general case where e > 2, it can be shown that the search for a lossless code need be a finite one only as follows:

The highest power of n in (1) is in the term $\frac{n^e}{e!}$. It is therefore, possible to rewrite (1) as follows:

$$n^e\ (1 + \varepsilon) = e!\ 2^j \qquad (1c)$$

in which, for any given e, the quantity ε can be made arbitrarily small for large values of n.

The difference between the number of even and odd correctors should be zero in a lossless code, and this condition can be expressed by the relation:

$$\sum_{k=0}^{k=e} \sum^{t} (-1)^t \frac{(n-r)!}{(n-r-k+t)!(k-t)!} \cdot \frac{r!}{(r-t)!\,t!} = 0 \tag{4}$$

which is obtained by subtracting (3) from (2), member by member.

The highest power terms in n and r in the expression above are:

$$\sum^{s} (-1)^t \frac{(n-r)^{e-t}\, r^t}{(e-t)!\,t!} = \frac{(n-2r)^e}{e!} \tag{5}$$

all other terms being of the form $n^a r^b$ where $a + b < e$. It is seen thus that (4) will be satisfied when n and r are related by an expression of the form:

$$n = 2r\,(1 + \gamma) \tag{6}$$

in which, for any given e, γ can be made arbitrarily small by making n and r sufficiently large.

It will be noted now that r can be factored algebraically out of (3). The terms multiplied by r which are of the form $\frac{(r-1)!}{(r-2s+1)!\,(2s-1)!}$ could be fractional, but each term will be an integer if multiplied by the highest common denominator of r and 2s-1, h.c.d. (r,2s-1). Therefore, if the lowest common multiplier of all h.c.d. (r,2s-1)'s is factored out of r, l.c.m. (all h.c.d. (r,2s-1)'s) = r', the multiplication by r' of all terms multiplied by r in the left member of (4) will be integers in all cases, and in order to satisfy (4), it should be possible to write r in the form:

$$r = 2^a r'' \qquad r'' \leqslant r' \tag{7}$$

It will be further noted that r', and hence r" also, have the upper bound:

$$r',\ r'' \leqslant \text{l.c.m. (all } (2s-1)\text{'s)},\ 2s-1 \leqslant e \tag{8}$$

Elimination of n and r between (1c), (6) and (7) gives:

$$2^{e(a+1)} r''^e (1+\gamma)^e (1+\varepsilon) = e!\, 2^j \tag{9}$$

For any given e, γ and ε approach zero for increasing j, while r" has an upper bound. Therefore an upper bound for j exists, beyond which (9) will not be satisfiable, because either e! will contain odd prime factors not contained in the left member of (9), or the left member of (9) will contain a number of odd prime factors which is a multiple of e, and which exceeds the number of the same odd prime factors in e!.

While this demonstration indicates that the search for lossless two or more symbol correcting binary codes need be a limited one only for any chosen number of errors, a search for such codes has only revealed, outside of the trivial cases of n errors in a 2n+1 binit message, the case mentioned earlier of a 3-error out of a 23-binit message symbol correcting code.. Whether, with the exception of the trivial cases mentioned, this particular 3-error symbol correcting code is the only lossless binary code correcting more than one symbol, is a matter of speculation. The degree of rarity of the happenstance required for the satisfaction of both relations (1) and (2) suggests that it could be so indeed, and offers the challenge of finding a mathematical demonstration of the impossibility to satisfy (1) and (2) for any other case.

Message Correcting Codes

The demonstration above leaves open the question of whether there are lossless e-error message correcting binary codes for any length of message, for condition (1) only applies to these, while condition (2) does not, since it is predicated upon the existence of a lossless symbol correcting code. For instance, the question is left open, whether a 2-error correcting 90 binit message code exists, since condition (1) is satisfied for this case.

The possibility that lossless message correcting codes exist where lossless symbol correcting codes do not is thus predicated upon the circumstance that message correcting codes form a more general class of codes. While no lossless binary message correcting codes are known, for which there are no corresponding lossless symbol correcting codes, examples will be given below of lossy message correcting codes which are more efficient than the available symbol correcting codes for the same number of

of message symbols and maximum allowable number of errors.

Some of these examples will be derived from the a_{mk} matrix already published in the referenced Letter to the Editor, and the formation of the top ten symbols in the Y_2 to Y_{12} columns will be explained briefly first.

If we consider five straight lines in a plane, A, B, C, D and E and order their respective intersections as follows: AB, AC, AD, AE, BC, BD, BE, CD, CE, DE, then Y_2 is formed by associating a 0 with the four intersections represented by the products in the expression:

$$A(B + C + D + E)$$

and a 1 with all other positions. Y_3, Y_4, Y_5, and Y_6 are formed likewise by associating a 0 with the intersections of B, C, D and E respectively with all four remaining lines.

Y_7 is formed by associating a 0 with the 5 intersections AB, BE, ED, DC, and CA of neighboring lines (including the first and last) in the operator:

$$(ABEDC)$$

which will be designated to represent the ensemble of 5 intersections listed above.

There are 4! cyclical permutation of the five lines, which can be separated into two groups of 12, the members of any one group being derivable from the other 11 members by an even number of interchanges of elements so that they can be said to be of the same parity. Within each group of 12 there are 6 pairs of permutations which differ only by their order, so that both members of each pair determine the same ensemble of 5 intersections. Thus, there will be only 6 district ensembles of 5 intersections having the same parity, and those belonging to the parity of the operator written above for Y_6 will determine the 0's of the upper 10 places of the Y_6 to Y_{12}.

It can be verified by inspection that the upper 10 symbols of Y_2 to Y_{12}, as well as the Boolean additions of any two of these ensembles, differ from all others in at least three places. Thus we can form the ensemble of 66 10-symbol messages written in Table I, which, together with the all 0's and all 1's messages form 68 10-symbol messages which differ in at least three places, and are therefore 1-error correcting messages.

TABLE I

0011100010100011110:1011100101101010:1011100101011001101001110101111
0111011000010100010:0101010011100110111:1000101001101111011111101011001
0101010100101010001:1100101100100111011:0111011001100011101010101111101
0100101001100100101:1011011001010111100:1101010111010100101011011001111
1000111010000100011:0110111011001001110:1101000011010111011101101101110
1010110101000110000:1001000110011101111:0100111100010101111111011110010
1001000110010101011:1100111001011000011:1010110110100110100111001101111
1100000001011011101:0011111010101001011:0011111010011010011101101010111
11111111111111111111:00000000000000000000:11111111111111111111000000000000
11111111111111111111:11111111111111111111:00000000000000000000000000000000

Likewise, the two smaller blocks of 36 9-symbol messages and 18 8-symbol messages, shown within the dashed enclosures, indicate that, together with the 2-all 0's and all 1's messages, there are 38-1 error correcting 9-symbol messages, and 20 1-error correcting 8-symbol messages.

On the other hand, it can be easily verified that there are only 16, 32 and 64 messages possible on the basis of 1 error correcting symbol codes for 8, 9 and 10-symbol messages respectively, because the number of cases of zero or one error are 9, 10 and 11 respectively in these three cases, which requires the assignment of 4 redundant symbols to the removal of the equivacation, thus leaving only 4, 5 and 6 symbols respectively for the message transmission.

It will also be noted that the upper 10 places of all Y_2 to Y_{12}, plus the all 0's message, form an ensemble of 12 10-symbol messages each of which differ from all others by at least five symbols, and are, therefore, 2-error correcting messages. The number of ways in which 0, 1 or 2 errors can occur in 10 places is: $1 + 10 + 45 = 55$, which indicates that a minimum of 6 redundant symbols should be assigned to the removal of the equivacation, thus leaving at most 4 symbols for the message transmission. However, it can be verified by inspection that it is impossible to form 4 6-symbol characteristics, which together with the 6 correctors for redundant X's constitute an ensemble in which any

member of which, and any sum of two of which differ from all other single members or sums of two.

On the other hand, it is possible to assign 7 symbols to the removal of the equivocation, and to have 3 7-symbol characteristics satisfying the conditions required so that only 3 symbols become thus available for the transmission of only 8 possible messages. Thus, here again, a larger number of messages can be transmitted by message coding than by symbol coding.

When the formation of 2-error correcting message codes is extended to 15-binit messages, in which the 15 intersections of 5 straight lines are associated in various ways with the message symbols, more care is required for the selection of favorable symetries.

Thus, we may associate the five intersections:

$$A(B + C + D + E + F)$$

with five 0's and the 6 groups of intersections of any one of the six lines with all others gives us 6 messages sufficiently distant from each other for 2-error correction.

We may consider next the 15 groups of intersections given by the various products of the form

$$(A + B)\ \ (C + D + E + F)$$

and we can verify that these vary in at least 5 symbol positions with each other and with those of the preceding form.

The 10 groups of intersections determined by expressions of the form

$$(A + B + C)\ \ (D + E + F)$$

can be added likewise to the other groups while satisfying the required criterion of a minimum of 5-symbol separations for 2-error correcting messages.

The 6 groups of 5 0's represented by the operator

$$(A\ B\ C\ D\ E)$$

and the five other operators of the same parity derivable from it can be verified to represent messages sufficiently distant from all others to permit 2-error correction. The letter F can be substituted for any and all other letters provided any other two letters are interchanged whenever a substitution is made, to provide more messages satisfying the 5 symbol distance criterion. Thus, 36 messages of this last type can be formed.

The total of messages satisfying the 5-symbol distance criterion which can be formed as indicated above is therefore:

$$6 + 15 + 10 + 36 = 67$$

It can be verified further that the 67 new messages formed by the boolean addition of the all 1's message to these satisfy the criterion with the 67 old messages. Adding the all 0's and all 1's messages gives us the total of 136 messages for the case of 2-error correcting 15-symbol messages.

Up to 2 errors can occur in a 15 symbol message in $1 + 5 + 105 = 121$ ways, and the upper bound to the number of theoretically possible is therefore $\frac{2^{15}}{121}$. The number of possible messages found above, 136, is seen to be slightly over half the number given by that upper bound. With a symbol-correcting code, 128 messages, i.e. slightly less than half the upper bound stated, could be transmitted by means of the a_{mk} matrix given in Table II.

TABLE II

Matrix for 2-error Symbol Correction of 15-symbol Messages

1	1	1	0	1	0	0
1	1	1	0	0	0	1
1	1	0	1	0	0	0
1	1	0	0	1	1	1
1	0	1	1	0	1	0
1	0	1	0	0	1	1
1	0	0	1	1	1	0
1	0	0	1	1	0	1

The examples of message coding given above suggest the question of whether the procedures described could be made methodical and be extended to longer messages. This question cannot be answered at this stage; instead circumstances will be pointed out which make such an answer difficult.

In the case just examined of a 15-symbol code in which the symbol positions were associated with the 15 intersections of 6 straight lines in a plane, a restricted number only of line groupings were studied. For instance, messages in which the 0's or 1's are given by the intersection of elements not above each other in the two lines of the matrix $\left|\begin{matrix}ABC\\DEF\end{matrix}\right|$ constitute another symetrical grouping, which examination indicated not to be useful in building 2-error correcting 15-symbol message codes, but which could be useful in other codes. Thus, a yet unsystematized selection of favorable groups must be made.

Codes may be based on the restricted class of n (2n-1) symbol messages ($n \leqslant 3$) which can be formed by assigning the symbols 0 or 1 to the positions determined by the intersections of 2n lines, A_1, A,....A_{2n}, given by all expressions of the form:

$$(A_1 + \ldots . A_{2m})\ (A_{2m+1} - \ldots + A_{2n})$$

and by assigning 1 versus 0 to all other points.

Together with the 2 all 0's and all 1's messages, these number $2^n - 1$ and are n-3 error correcting. Thus, 15, 28 and 45 symbol messages will be in number 2^5, 2^7, and 2^9 and will be 3, 5 and 7 error correcting respectively. This code equals the Reed code in the case of 15-symbol messages, and is inferior to it for longer messages. A short examination of codes formed by considering the intersections of the planes A_1, A_2, - - - - A_n, in a 3 dimensional space which are of the form:

$$(A_1 + - - + A_k)\ (A_{k+1} + - - + A_m)\ (A_{m+1} + - - + A_n)$$

has not indicated that an extension of this attack to multidimensional spaces is promising. There again, a selection of proper symetries is required.

Another circumstance to be pointed out here is the completely symetrical part played by all straight lines in the formation of the 1 or 2-error correcting 10-symbol messages and 2-error correcting 15-symbol messages described above. By contrast, an examination of the lossless 1-error symbol correcting 15 symbol message code can be seen to be expressible in terms of the intersections of 6 lines in which 4 lines play symetrical roles, but the other 2 do not. This may permit the speculation that an approach to the problems of building a 2-error correcting 90 symbol message codes of 2^{78} messages may be to consider the 90 intersections of 14 lines in which 2 lines, the intersection of which is not counted, play a part not symetrical with that played by the 12 others.

CONCLUSION

It has been shown that lossless[e] symbol correcting message codes can exist only for message lengths which have an upper bound, and it can be speculated whether any exist, outside of the cases of 1-error correcting 2^n-1 symbol messages, n-error correcting 2n + 1 symbol messages, and 3-error correcting 23-symbol messages.

It has also been shown by examples that the more general class of message correcting codes permits a higher coding efficiency than symbol correcting codes, and the existence of lossless message correcting codes not included within the lossless symbol correcting codes mentioned above appears less improbable.

While the only systematic message correcting codes described in the text is less efficient than the Reed Code, it is suggested that an attack along these lines may prove more fruitful than if restricted to the sub-class of symbol correcting codes, when attempts are made to design systematic codes approaching Shannon's upper bound.

ERROR-FREE CODING*

Peter Elias

Department of Electrical Engineering and Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Introduction

This paper describes constructive procedures for encoding messages to be sent over noisy channels so that they may be decoded with an arbitrarily low error rate. The procedures are a kind of iteration of simple error-correcting codes such as those of Hamming[1] and Golay[2]; any additional systematic codes which may be discovered, such as those discussed by Reed[3] and Muller[4], may be iterated in the same way.

The procedures are not ideal; that is, the capacity of a noisy channel for the transmission of error-free information using such coding is smaller than information theory says it should be. However, the procedures do permit the transmission of error-free information at a positive rate. They also have these two properties.

(1) The codes are "systematic" in Hamming's sense: they are what Golay calls "digit codes" rather than "message codes." That is, the transmitted symbols are divided into so-called "information digits" and "check digits." The customer who has a message to send supplies the information digits which are transmitted unchanged. Periodically the coder at the transmitter computes some check digits, which are functions of past information digits, and transmits them. The customer with a short message does not have to wait for a long block of symbols to accumulate before coding can proceed, as in the case of codebook coding, nor does the coder need a codebook memory containing all possible symbol sequences. The coder needs only a memory of the past information digits it has transmitted and a quite simple computer.

(2) The error probability of the received messages is as low as the receiver cares to make it. If the coding process has been properly selected for a given noisy channel, the customer at the receiver can set the probability of error per decoded symbol (or the probability of error for the entire sequence of decoded symbols transmitted up to the present, or the equivocation of part or all of the decoded symbol sequence) at as low a value as he chooses. It will cost him more delay to get a more reliable message, but it will not be necessary to alter the coding and decoding procedure when he raises his standards, nor will it be necessary for less particular and more impatient customers using the same channel to put up with the additional delay. This is again unlike codebook processes, in which the codebook must be rewritten for all customers if any one of them raises his standards.

Perhaps the simplest way to indicate the basic behavior of such codes is to describe how one would work in a commercial telegraph system. A customer entering the telegraph office presents a sequence of symbols which are sent out immediately over a noisy channel to another office, which immediately reproduces the sequence, adds a note "the probability of error per symbol is 10^{-1}, but wait till tomorrow," and sends it off to the recipient. Next day the recipient receives a note saying "For 'sex' read 'six'. The probability of error per symbol is now 10^{-2}, but wait till next week." A week later the recipient gets another note: "For 'lather' read 'gather'. The probability of error per symbol is now 10^{-4}, but wait till next April." This flow of notes continues, the error probability dropping rapidly from note to note, until the recipient gets tired of the whole business and tells the telegraph company to stop bothering him.

Since these coding procedures are derived by an iteration of simple error-correcting and detecting codes, their performance depends on what kind of code is iterated. For a binary channel with a small and symmetric error probability, the best choice among the available procedures is the Hamming-Golay single-error-correction double-error-detection code developed by Hamming[1] for the binary case and extended by Golay[2] to the case of symbols selected from an alphabet of M different symbols, where M is any prime number. The analysis of the binary case will be presented in some detail and will be followed by some notes on diverse modifications and generalizations.

---

*This work was supported in part by the Signal Corps; the Office of Scientific Research, Air Research and Development Command; and the Office of Naval Research.

Iterated Hamming Codes

First-Order Check

Consider a noisy binary channel, which transmits each second either a zero or a one, with a probability $(1 - p_o)$ that the symbol will be received as transmitted, and a probability p_o that it will be received in error. Error probabilities for successive symbols are assumed to be statistically independent.

Let the receiver divide the received symbol sequence into consecutive blocks, each block consisting of N_1 consecutive symbols. Because of the assumed independence of successive transmission errors, the error distribution in the blocks will be binomial: there will be a probability

$$P(o) = (1 - p_o)^{N_1}$$

that no errors have occurred in a block, and a probability P(i)

$$P(i) = \frac{N_1!}{i!\,(N_1 - i)!}\, p_o^i (1 - p_o)^{N_1 - i} \tag{1}$$

that exactly i errors have occurred.

If the expected number of errors per received block, $N_1 p_o$, is small, then the use of a Hamming error-correction code will produce an average number of errors per block, $N_1 p_1$, after error correction, which is smaller still. Thus p_1, the average probability of error per position after error correction, will be less than p_o. An exact computation of the extent of this reduction is complicated, but some inequalities are easily obtained.

The single-error-correction check digits of the Hamming code give the location of any single error within the block of N_1 digits, permitting it to be corrected. If more errors have occurred, they give a location which is usually not that of an incorrect digit, so that altering the digit in that location will usually cause one new error, and cannot cause more than one. The double-error-detection check digit tells the receiver whether an even or an odd number of errors has occurred. If an even number has occurred and an error location is indicated, the receiver does not make the indicated correction, and thus avoids what is very probably the addition of a new error.

The single-correction double-detection code, therefore, will leave error-free blocks alone, will correct single errors, will not alter the number of errors when it is even, and may increase the number by at most one when it is odd and greater than one. This gives for the expected number of errors per block after checking

$$\begin{aligned} N_1 p_1 &\leq \sum_{\text{even } i \geq 2}^{\leq N_1} i\,P(i) + \sum_{\text{odd } i \geq 3}^{\leq N_1} (i+1)\,P(i) \\ &\leq P(2) + \sum_{i=3}^{N_1} (i+1)\,P(i) \\ &\leq \sum_{i=0}^{N} (i+1)\,P(i) - P(0) - 2P(1) - P(2) \\ &\leq 1 + N_1 p_o - P(0) - 2P(1) - P(2). \end{aligned} \tag{2}$$

Substituting the binomial error probabilities from (1), expanding and collecting terms, gives, for $N_1 p_o \leq 3$,

$$\begin{aligned} N_1 p_1 &\leq N_1(N_1 - 1)\, p_o^2, \\ p_1 &\leq (N_1 - 1)\, p_o^2 < N_1 p_o^2. \end{aligned} \tag{3}$$

The error probability per position can therefore be reduced by making N_1 sufficiently small. The shortest code of this type requires $N_1 = 4$, and the inequality (3) suggests that a reduction will therefore not be possible if $p_o \geq 1/3$. The fault is in the equation, however, and not the code: for $N_1 = 4$ it is a simple majority-rule code which will always produce an improvement for any $p_o < 1/2$.

A Hamming single-correction double-detection code uses C of the N positions in a block for checking purposes and the remaining N - C positions for the customer's symbols, where

$$C = \left[\log_2 (N-1) + 2\right]. \tag{4}$$

(Here and later, square brackets around a number denote the largest integer which is less than or equal to the number enclosed. Logarithms will be taken to the base 2 unless otherwise specified.)

Higher Order Checks

After completing the first-order check, the receiver discards the C_1 check digits, leaving only the $N_1 - C_1$ checked information digits, with the reduced error probability p_1 per position. (It can be shown that the error probability after checking is the same for all N_1 positions in the block, so that discarding the check digits does not alter the error probability per position for the information digits.) Now some of these checked digits are made use of for further checking, again with a Hamming code. The receiver divides the checked digits into blocks of N_2; the C_2 checked check digits in each block enable it, again, to correct any single error in the block, although multiple errors may be increased by one in number. In order for the checking to reduce the expected number of errors per second-order block, however, it is necessary to select the locations of the N_2 symbols in the block with some care.

The simplest choice would be to take several consecutive first-order blocks of $N_1 - C_1$ adjacent checked information digits as a second-order block, but this is guaranteed not to work. For if there are any errors at all left in this group of digits after the first-order checking, there are certainly two or more, and the second-order check cannot correct them. In order for the error probability per place after the second-order check to satisfy the analog of (3), namely,

$$p_j \leq (N_j - 1)\, p_{j-1}^2 < N_j\, p_{j-1}^2, \tag{5}$$

it is necessary for the N_2 positions included in the second-order check to have statistically independent errors after the first check has been completed. This will be true if, and only if, each position was in a different block of N_1 adjacent symbols for the first-order check.

The simplest way to guarantee this independence is to put each group of $N_1 \times N_2$ successive symbols in a rectangular array, checking each row of N_1 symbols by means of C_1 check digits, and then checking each column of already checked symbols by means of C_2 check digits. The procedure is illustrated in Fig. 1. The transmitter sends the $N_1 - C_1$ information digits in the first row, computes the C_1 check digits and sends them, and proceeds to the next row. This process continues down through row $N_2 - C_2$. Then the transmitter computes the C_2 check digits for each column and writes them down in the last C_2 rows. It transmits one row at a time, using the first $N_1 - C_1$ of the positions in that row for the second-order check, and the last C_1 digits in the row for a first-order check of the second-order check digits.

After the second-order check, then, the inequality (5) applies as before, and we have for p_2, the probability of error per position,

$$p_2 < N_2\, p_1^2 < N_2\, N_1^2\, p_o^4. \tag{6}$$

The N_3 digits to be checked by the third-order check may be taken from corresponding positions in each of N_3 different $N_1 \times N_2$ rectangles, the N_4 digits in a fourth-order block from corresponding positions in N_4 such collections of $N_1 \times N_2 \times N_3$ symbols each, and so on ad infinitum. At the k^{th} stage this gives

$$p_k < N_k^{2^0} \cdot N_{k-1}^{2^1} \cdots N_{k-j}^{2^j} \cdots N_1^{2^{k-1}} \cdot p_o^{2^k}. \tag{7}$$

It is now necessary to show that not all of the channel is occupied, in the limit, with checking digits

of one order or another so that some information can also get through. The fraction of symbols used for information at the first stage is $\left[1 - (C_1/N_1)\right]$. At the k^{th} stage, it is

$$F_k = \prod_1^k \left(1 - \frac{C_j}{N_j}\right). \tag{8}$$

It is now necessary to find a sequence of N_j for which p_k approaches zero and F_k does not, as k increases without bound. A convenient sequence is

$$N_1 = 2^n$$

$$N_j = 2^{j-1} N_1 = 2^{j+n-1}. \tag{9}$$

This gives for p_k, from (7),

$$p_k < \left(N_1 \cdot 2^{k-1}\right)^{2^0} \cdots \left(N_1 \cdot 2^{k-j}\right)^{2^{j-1}} \cdots \left(N_1 \cdot 2^0\right)^{2^{k-1}} p_o^{2^k}$$

$$< \frac{1}{N_1} (2N_1 p_o)^{2^k} \cdot 2^{-(k+1)}. \tag{10}$$

The right side of (10) approaches zero as k increases, for any $N_1 p_o \leqslant 1/2$. Thus the error probability can be made to vanish in the limit. Note that the inequality gives a much weaker kind of approach to zero for the threshold value $N_1 p_o = 1/2$ than for any smaller value of errors per first-order block.

For the same sequence of N_j, a lower bound on F_∞ can be computed. From equations (8) and (4) we have

$$F_\infty = \prod_1^\infty \left(1 - \frac{C_j}{N_j}\right) = \prod_1^\infty \left(1 - \frac{\log_2 N_j + 1}{N_j}\right)$$

$$= \prod_1^\infty \left(1 - \frac{j+n}{2^{j+n-1}}\right). \tag{11}$$

Let

$$\sigma_j = \frac{C_j}{N_j}$$

$$\sigma = \sum_1^\infty \sigma_j. \tag{12}$$

Then σ_j is monotonic decreasing in j and is less than 1 for all constructable Hamming codes, that is, for $N_1 = 2^n \geqslant 4$. This makes it possible to write the following inequalities:

$$e^{-\sigma} > F_\infty > (1 - \sigma_1)^{\sigma/\sigma_1} > 1 - \sigma. \tag{13}$$

Here the last term on the right is one of the Weierstrasse inequalities for an infinite product; the other terms are useful when $\sigma > 1$, and show that for $\sigma_1 < 1$ and $\sigma < \infty$, F_∞ is strictly positive.

Evaluating σ in the present case gives

$$\sigma = \sum_{j=1}^\infty \frac{j+n}{2^{j+n-1}} = \frac{n+2}{2^{n-1}} = \frac{2 \log 4N_1}{N_1}. \tag{14}$$

At threshold, that is, at $N_1 p_o = 1/2$, this gives

$$\sigma = 4 p_o \log \frac{2}{p_o} < 4 \left\{ p_o \log \frac{1}{p_o} + (1 - p_o) \log \frac{1}{1 - p_o} \right\} = 4E \tag{15}$$

where E is the equivocation of the noisy channel. Thus for p_o small, from (13) we have

$$F_\infty > 1 - 4E. \tag{16}$$

That is, under the specified conditions ($N_1 p_o = 1/2$, $N_1 = 2^n \geq 4$) the number of check digits required is never more than four times the number that would be required for an ideal code, provided that an ideal code of the check-digit type exists, which is not obvious. When E is > 1/4, the interior inequality shows that F_∞ is still positive.

Equivocation

Feinstein[3] has shown that it is possible to find ideal codes for which not only the probability of error, but the total equivocation, vanishes in the limit as longer and longer symbol sequences are used. This property is also true for the coding processes described here. This is a very important result in the case of codebook codes, where the message becomes infinite in the limit. For the codes under discussion here, it is a less important property, since any finite message can be received without an infinite lag, and its equivocation vanishes with the error probability per position.

The total number of binary digits checked by the k^{th} checking stage is

$$M_k = \prod_1^k N_j . \tag{17}$$

Of these $F_k M_k$ are information digits and the remainder are checks. Using the values (9) for the N_j, we have

$$M_k = N_1^k 2^{\frac{k(k-1)}{2}} . \tag{18}$$

The bound (10) limits the probability of error per position. Multiplying this by M_k gives a bound on the mean number of errors per M_k digits, which is also a bound on the fraction of sequences of M_k digits which are in error after checking – a gross bound, since actually any such sequence which is in error must have many errors, and not just one. Thus for Q_k, the probability that a checked group of M_k digits is in error, we have

$$Q_k < p_k M_k \leq \frac{1}{4} \left(\frac{N_1}{2} \right)^{k-1} (2N_1 p_o)^{2^k} 2^{\frac{k(k-1)}{2}} . \tag{19}$$

At threshold ($N_1 p_o = 1/2$) this inequality does not guarantee convergence, but for $N_1 p_o < 1/2$, Q_k certainly approaches zero as k increases.

The equivocation E_k per sequence of M_k terms is bounded by the value it would have if any error in a block made all possible symbol sequences equally likely at the receiver, that is,

$$E_k < Q_k \log \frac{1}{Q_k} + (1 - Q_k) \log \frac{1}{1 - Q_k} + Q_k M_k . \tag{20}$$

Again at threshold convergence is not guaranteed, but for $N_1 p_o < 1/2$, E_k, the absolute equivocation of the block, will also vanish as k increases.

Distance Properties

At the k^{th} stage of this coding process, a sequence of M_k binary digits has been selected as a message. Because the check digit values are determined by the information digit values, there are only $2^{F_k M_k}$ possible message sequences, rather than 2^{M_k}. Any two of these possible messages will have a

"distance" from one another, defined as the number of positions in which they have different binary symbols, and the smallest such distance will be 4^k for the iterated single-correction, double-detection code. This means that by using this set of codes with a codebook, any set of errors less than one-half of the minimum distance in number can be corrected by choosing as the transmitted message the message point nearest to the received sequence.

It is easy to see that for the coding procedure just described this error-correction capability will not be realized. Any set of 2^k errors which are at the corners of a k-dimensional cube in the k-dimensional rectangle of symbol positions will not be corrected by this process, since each check will merely indicate a double error which it cannot correct. By inspecting any two of the sets of check digits at once, these errors could be located, but they will not have been corrected by the process as described above. The effective minimum of the maximum number of errors which will be corrected is therefore $2^k - 1$, rather than $2^{2k-1} - 1$.

This shows a loss of error-correction capability because of the strictly sequential use of the checking information. Without going to the extreme memory requirement of codebook techniques, a portion of this loss may be recouped by not throwing the low-order check digits away but using them to recheck after higher order checking has been done. This does not increase the maximum number of errors for which correction is always guaranteed, but it does reduce the average error probability at each stage; the exact amount of this reduction is, unfortunately, difficult to compute. This behavior, however, points up a significant feature of the coding process. If the maximum number of errors for which correction is always guaranteed were the maximum number of errors for which correction was ever guaranteed, the procedure could not transmit information at a nonzero rate; that is, the minimum distance properties of the code are inadequate for the job. It is average error-correction capability that makes transmission at a nonzero rate possible.

The Poisson Limit

Much of the above analysis has assumed that $N_j = 2^{j-1} N_1$, and part of it has further assumed that $N_1 = 2^n$. However, any series of N_j which increases rapidly enough so that σ is finite will lead to a coding process that is error-free for sufficiently small values of $N_1 p_o$. In particular, any other approximately geometric series may be used, for which

$$N_j \approx b^{j-1} N_1, \quad b > 1. \tag{21}$$

The approximation is necessary if b is not an integer. The expression for p_k analogous to (10) is then

$$p_k < \frac{1}{N_1} (bN_1 p_o)^{2^k} b^{-(k+1)}, \tag{22}$$

with a threshold at $N_1 p_o = 1/b$. The value of σ can also be bounded for this series. At threshold, the bound corresponding to (15) is

$$\sigma \leq \frac{b^2 p_o}{b-1} \log\left(\frac{4}{p_o \, b \, \frac{b-2}{b-1}}\right). \tag{23}$$

Again, for $N_1 p_o$ below threshold, Q_k and E_k approach zero as k increases.

For very small p_o, the value of b that minimizes σ is b = 2. This leads to the maximum value of F_∞ given by (16). However, for very small p_o, N_1 may be made very large. The distribution of errors in the blocks then approaches the Poisson distribution, for which the probability that just i errors have occurred in a block is

$$P(i) = e^{-N_1 p_o} \cdot \frac{(N_1 p_o)^i}{i!}. \tag{24}$$

This equation may be used to derive an iterative inequality on the mean number of errors per block after single-detection, double-correction coding.

$$\begin{aligned} N_j p_j &\leq 1 + N_j p_{j-1} - e^{-N_j p_{j-1}} \left\{1 + 2N_j p_{j-1} + \frac{(N_j p_{j-1})^2}{2!} + \frac{(N_j p_{j-1})^4}{4!} + \ldots\right\} \\ &\leq 1 - N_j p_{j-1} \left(2e^{-N_j p_{j-1}} - 1\right) - \frac{1}{2}\left(1 - e^{-2N_j p_{j-1}}\right). \end{aligned} \tag{25}$$

Keeping $N_j p_j$ constant gives the geometric series (21) for N_j. A joint selection of $N_1 p_o$ and b for the minimization of the bound on σ gives $N_1 p_o \approx 0.75$, $b \approx 1.75$, and an effective channel capacity

$$F_\infty \sim 1 - 3.11E, \tag{26}$$

where E is the equivocation of the binary channel. This is an improvement over (16).

Iteration of Other Codes

The analysis in the preceding sections has dealt only with iteration of the Hamming single-error-correction, double-error-detection code. Other kinds of codes may also be iterated; nor is it necessary to use the same type of code at each stage in the iterative process. The only requirement is that each code be of the check-digit, or systematic, type, so that its check digits may be computed on the basis of the preceding information digits and added on to the message.

First, the final parity check digit of a Hamming code may be omitted, destroying the double-detection feature of the code. This leads to the inequality

$$p_j \leqslant \frac{3}{2}(N_j - 1)\, p_{j-1}^2 < \frac{3}{2} N_j\, p_{j-1}^2, \tag{27}$$

in place of (5). Iterating this code alone gives a bound on σ that is only slightly smaller than (15), but the threshold becomes $N_1 p_o = 1/3$ rather than $N_1 p_o = 1/2$, and the effective channel capacity for small p_o is bounded by

$$F_\infty > 1 - 6E, \tag{28}$$

where E is the equivocation of the binary channel.

Second, the Golay[2] analogs to both kinds of Hamming code may be constructed, for M-ary channels, where M is a prime number. If there is a probability $(1 - p_o)$ that any symbol will be received correctly, and if the consecutive errors are statistically independent, the results of the binary case carry over quite directly. The inequalities (5) and (27) still hold for the two kinds of codes, since the errors as a whole are still binomially distributed in blocks. At threshold, inequalities (16) and (28) still hold for the effective channel capacity, where E is now the equivocation of a symmetrical M-ary channel; that is, of a channel in which the probability of an error taking any given symbol into any other different symbol is $p_o/(M-1)$. The result (26) for the Poisson limit also applies, with the same interpretation of E.

Third, the Reed[4]-Muller[5] codes may be treated as check digit codes, and may be iterated to give an error-proof system. For these codes, the average error-reduction capability is not known; only the minimum distance is known. Certain of the codes, such as the triple-correction quadrupole-detection code for blocks of 32 binary symbols, might provide a good starting point for an iteration which proceeds by iteration of Hamming codes. The Golay triple-correction quadruple-detection code for blocks of 24 symbols might be used in the same way. It will take considerable computation to evaluate such mixed iteration schemes.

It is not, at present, profitable to use the Reed-Muller codes for later stages in the iteration. The reason is that an efficient triple-correction quadruple-detection code should require about $C = 2 \log N$ check digits for a block of length N. The Reed-Muller codes require about $C = 1 + \log N + 1/2 \log N (\log N - 1)$ check digits for this purpose. For large N, therefore, the effective channel capacity is reduced by the large number of check digits required. There is a similar inefficiency in the Reed-Muller codes with greater error-correction capabilities, which might be removed if the average error-correction capabilities of these codes were known.

Nonrectangular Iteration

The problem of assuring statistical independence among the N_k digits checked by a k^{th} order check, so that the inequality (5) derived on the basis of statistical independence can be used as an iterative inequality, was solved above by what might be called rectangular iteration. Each of the N_k digit positions in a check group are selected from a different sequence of M_{k-1} consecutive symbols. Thus until the k^{th} order checking has been carried out, no two of them have been associated by lower order checking procedures in any way. This iteration solves the problem, but it makes M_k a function that grows very rapidly with k. When N_j is the geometric series (21), then

$$M_k \approx N_1^k\, b^{\frac{1}{2}k(k-1)} \tag{29}$$

This means that p_k, Q_k, and E_k decrease quite rapidly as functions of k, but much more slowly as functions of the length of the message M_k, or its information content $F_k M_k$.

Roughly speaking, if $H_k = F_k M_k$ is the total number of information digits transmitted at the k^{th} stage,

$$p_k \sim A e^{-2^{a(\log H_k)^{1/2}}}, \quad A > 0, \quad a > 0. \tag{30}$$

This is a much slower decrease of error probability than Feinstein's result[3] which is

$$p_k \sim B e^{-b \cdot 2^{\log H_k}} = B e^{-bH_k}, \quad B > 0, \ b > 0. \tag{31}$$

A less stringent requirement on the choice of digits checked in a single group is that no two of them have been together in any lower order check group. This requires that there be at least N_k different groups of order k - 1 from which to select digit positions. Thus

$$M_k \geqslant N_k N_{k-1}. \tag{32}$$

If it is possible to approximate equality in (32), and if the statistical dependence so introduced does not seriously weaken inequality (5), then it might be possible to get the result

$$p_k \sim D e^{-2^{d \log H_k}}, \quad D > 0, \ d > 0, \tag{33}$$

which is closer to Feinstein's result.

Conclusion

From a practical point of view, this coding procedure has much to recommend it. A question of both theoretical and practical interest is the extent to which the convenience associated with a computable and error-free code is compatible with ideal coding, or the smallest price that must be paid for the convenience if the two are incompatible. No answer to this question is in sight at present. However, the existence of the error-free process, despite its lack of ideality, puts the burden of efficient coding on the first stage of the coding process. For if a coding process succeeds in reducing the equivocation in a received message to some small but positive value E, the remaining errors may always be eliminated at a cost of 4 E (or 3.11 E) in channel capacity: an error-proof termination is available, at a price, to take care of the residual errors left by any other error-correcting scheme.

Acknowledgment

The iterative approach used in this paper was suggested by a comment of Dr. Victor H. Yngve, of the Research Laboratory of Electronics, M.I.T., on the fact that redundancy in language was added at many different levels, a point that he discusses in reference 6.

References

(1) R. W. Hamming, Error Detecting and Error Correcting Codes, Bell System Tech. J. 29, pp. 147-160 (1950).

(2) M. J. E. Golay, Notes on Digital Coding, Proc. I.R.E. 37, p. 657 (1949).

(3) A. Feinstein, Some New Basic Results in Information Theory, these transactions.

(4) I. S. Reed, A Class of Multiple-Error-Correcting Codes and the Decoding Scheme, Technical Report No. 44, Lincoln Laboratory, M.I.T., (1953). See also the paper under this title in these transactions.

(5) D. E. Muller, "Metric Properties of Boolean Algebra and their Applications to Switching Circuits," Report No. 46, Digital Computer Laboratory, University of Illinois (1953).

(6) V. H. Yngve, "Language as an Error-Correcting Code," pp. 73-74, Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., April 15, 1954.

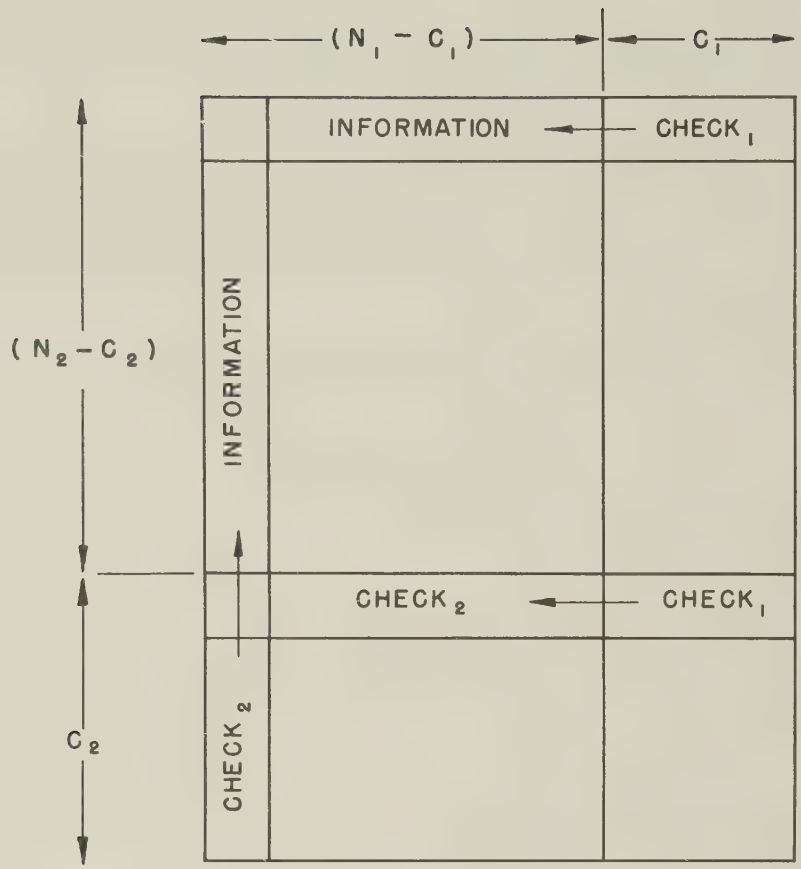


Fig. 1 – Organization of First- and Second-Order Check Digits.

A CLASS OF MULTIPLE-ERROR-CORRECTING CODES AND THE DECODING SCHEME

Irving S. Reed

Lincoln Laboratory - Massachusetts Institute of Technology
Cambridge, Massachusetts

I. Introduction

A procedure for constructing one-error-correcting and two-error-detecting systematic codes was introduced in a recent study by R. W. Hamming.[1] It is the purpose of this paper to exhibit some examples of n-error-correcting and (n + 1) error-detecting systematic codes for the cases where both the code length and (n + 1) are powers of two. The class of codes to be considered was developed by D. E. Muller in his recent work.[2]

The decoding scheme presented in this paper differs from Hamming's scheme in that the encoded message will be extracted directly from the possibly corrupted received code by a majority testing of the redundant relations within the code. Hamming's scheme for n = 1 was dependent first on the location of a possible digit error in the code; secondly, on the correction of that digit; and lastly, on the extraction of the message from the corrected code. By circumventing Hamming's step of error location and correction, which is quite a severe problem when n is not equal to one, we have arrived at a decoding scheme that makes a natural use of the redundancy within the code as well as being conceptually simple.

In this paper, some of the mathematical proofs of the methods discussed will be avoided for the sake of brevity of exposition. A more detailed mathematical analysis will appear elsewhere.

II. Some Mathematical Preliminaries

A code having n binary digits may be considered the element of a space, consisting of 2^n elements of the form

$$f = (f_0, \ldots f_{n-1})$$

where

$$(f_j = 0, 1) \quad \text{for} \quad (j = 0, 1, 2, \ldots n-1) \ .$$

This space is technically an Abelian group if the sum of any two elements f and g in the space is defined as follows:

$$f \oplus g = (f_0, f_1, \ldots f_{n-1}) \oplus (g_0, g_1, \ldots g_{n-1}) = (f_0 \oplus g_0, f_1 \oplus g_1, \ldots f_{n-1} \oplus g_{n-1}) \ ,$$

where $f_j \oplus g_j$ is the sum modulo two of the binary digits f_j and g_j for $(j = 0,1,2, \ldots n-1)$. If multiplication by the binary scalar α is allowed as

$$\alpha f = \alpha(f_0, f_1, \ldots f_{n-1}) = (\alpha f_0, \alpha f_1, \ldots \alpha f_{n-1}) \ ,$$

the Abelian group may be termed a generalized vector space of n-dimensions or a module. Finally, if the product operation

$$f \cdot g = (f_0, f_1, \ldots f_{n-1}) \cdot (g_0, g_1, \ldots g_{n-1}) = (f_0 g_0, f_1 g_1, \ldots f_{n-1} g_{n-1})$$

for f and g in the module is introduced, the space is a Boolean ring. The prime operation is defined to be

$$f' = f \oplus I$$

for f in the ring, and where I is the identity vector $(1, 1, 1, \ldots 1)$.

Into this space one may further introduce a norm or length of a vector as follows:

$$\| f \| = \sum_{i=0}^{n-1} f_i$$

where Σ refers to ordinary addition. It is not difficult to see that the norm of the sum of two elements f and g in the ring or $\| f \oplus g \|$ is precisely the Hamming distance D(f,g) as defined in Ref. 1.

Now let n the dimension of the vector space be a power of two or $n = 2^m$. Let a vector of this space be of the form

$$f = (f_0, f_1, \ldots f_{2^m-1}) \ ,$$

where f_j is a binary digit for $(j = 0, 1, \ldots 2^m-1)$. Now the vector f may be clearly expressed as

$$f = f_0 I_0 \oplus f_1 I_1 \oplus \ldots f_{2^m-1} I_{2^m-1} \ , \tag{1}$$

where I_j is a unit vector with the digit one in j-th coordinate of the vector and zeros elsewhere for $(j = 0,1, \ldots 2^m-1)$. Further, each unit vector I_j can be determined as a product of m vectors from the set of 2m vectors $x_1, x_2, x_3, \ldots x_m, x_1', x_2', x_3', \ldots x_m'$, where x_1 is a vector consisting of alternating zeros and ones, beginning with zero; x_2 is a vector consisting of alternating zero pairs and one pairs, beginning with a zero pair, and so forth, as follows:

$$\begin{aligned}
x_1 &= (0\ 1\ 0\ 1\ 0\ 1\ 0\ 1 \ldots 0\ 1) \ , \\
x_2 &= (0\ 0\ 1\ 1\ 0\ 0\ 1\ 1 \ldots 1\ 1) \ , \\
x_3 &= (0\ 0\ 0\ 0\ 1\ 1\ 1\ 1 \ldots 1\ 1) \ , \\
&\ \ \vdots \\
x_m &= (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \ldots 1\ 1) \ .
\end{aligned} \tag{2}$$

If $x_k^{i_k}$ is defined to be x_k' for $i_k = 0$ and x_k for $i_k = 1$, then by the rules of Boolean algebra,

$$I_j = x_1^{i_1} x_2^{i_2} \ldots x_m^{i_m} \ , \tag{3}$$

where

$$j = \sum_{k=1}^{m} i_k 2^{k-1} \quad \text{with } (i_k = 0,1) \text{ for } (j = 0,1, \ldots m-1) \ .$$

Combining Eqs. (1) and (3), we have

$$f = \bigoplus_{j=0}^{2^m-1} f_j x_1^{i_1} x_2^{i_2} \ldots x_m^{i_m} \ , \tag{4}$$

where $i_1, i_2, \ldots i_m$ are the digits of the binary representation of j, and where the summation sign $\bigoplus$ is with respect to the sum operation $\oplus$. Equation (4) is the canonical expansion of any vector f in the Boolean algebra of 2^m dimensional vectors, consisting of binary digits.

If the identity $x_j' = I \oplus x_j$ and the distributive law of algebra is used, Eq.(4) may be expanded to obtain the following polynomial in the x_j's:

$$\begin{aligned}
f = g_0 \oplus g_1 x_1 \oplus \ldots \oplus g_m x_m \oplus g_{12} x_1 x_2 \oplus \ldots g_{m-1,m} x_{m-1} x_m \oplus \ldots \\
\ldots \oplus g_{12 \ldots m} x_1 x_2 \ldots x_m \ .
\end{aligned} \tag{5}$$

Equation (5) can be written more explicitly as

$$f = f(0,\ldots 0) \oplus \underset{1}{\Delta} f(0,\ldots 0)x_1 \oplus \ldots \oplus \underset{m}{\Delta} f(0,\ldots 0)x_m \oplus \underset{12}{\overset{2}{\Delta}} f(0,\ldots 0)\, x_1 x_2$$

$$\oplus \ldots \oplus \underset{12\ldots m}{\overset{m}{\Delta}} f(0,\ldots 0)x_1 x_2 \ldots x_m \quad , \tag{6}$$

where

$$f(i_1,\ldots i_m) = f_j \quad \text{when} \quad j = \sum_{k=1}^{m} i_k 2^{k-1} \quad \text{for} \quad i_k = 0,1 \quad ,$$

and the Δ's are multiple partial differences, for example,

$$\underset{1}{\Delta} f(0) = f(1,0,0,\ldots) \oplus f(0,0,0,\ldots 0) \quad ,$$

$$\underset{12}{\overset{2}{\Delta}} f(0) = \left[f(1,1,0,\ldots) \oplus f(0,1,0,\ldots)\right] \oplus \left[f(1,0,0,\ldots) \oplus f(0,0,0,\ldots)\right] \quad ,$$

and so forth. The polynomial representation in Eq.(6) of the vector f supplies the relations between the coefficients of Eq.(5) and the scalars f_j of Eq.(4) for $(j = 0,1,2,\ldots 2^m-1)$. This definition of the Δ's will be expanded in another section of this paper.

III. The Generation of the Multiple Error Allowing Codes

Suppose that the dimension of the space considered in the previous section is 2^m. Consider the set Φ_r^m of all polynomials of the form (5) of degree less than or equal to r where $r \leq m$. Each such polynomial must have the form

$$g_0 \oplus g_1 x_1 \oplus \ldots \oplus g_m x_m \oplus \ldots \oplus g_{12\ldots r}\, x_1 \ldots x_r \oplus \ldots \oplus g_{m-r+1,\ldots m}$$

$$x_{m-r+1} \ldots x_m \quad , \tag{7}$$

and the sum of any two such polynomials is a member of the same set. This implies that Φ_r^m the set of all polynomials of type (7) or of degree less than or equal to r forms an Abelian group or submodule of the Boolean ring of 2^m dimensional vectors. Since Φ_r^m is a module, the Hamming distance between any two elements of Φ_r^m is the norm of a third element of Φ_r^m . This fact was exploited by D. E. Muller[2] in proving his Theorem 25. Muller's Theorem 25, in our terminology, may be expressed as follows:

Theorem A:- The norms of all non-zero vectors f of Φ_r^m satisfy

$$\| f \| \geq 2^{m-r} \quad \text{for } (m = 0,1,2, \ldots) \quad \text{and } r \leq m \quad .$$

We shall not prove this theorem here. It suffices to say that Muller proved the theorem by an induction on m and r and the properties of the Hamming distance.

By the above theorem there is at least a distance 2^{m-r} between two elements of Φ_r^m and, as a consequence, there is an open Hamming sphere of radius 2^{m-r-1} about each element of Φ_r^m in Φ_m^m (the whole vector space) which does not intersect any other such sphere. This means that it is possible to associate each element of such a sphere with the element defining the sphere or what is the same to associate an element of Φ_r^m which is less than a distance 2^{m-r-1} from an element f of Φ_r^m with f .

In order to illustrate how a message may be coded into an error-detecting code of the type described above, consider the following example: Let $m = 4$ and $r = 1$, by (7) the vectors of Φ_1^4 are of the form

$$g_0 \oplus g_1 x_1 \oplus g_2 x_2 \oplus g_3 x_3 \oplus g_4 x_4 \quad . \tag{8}$$

Let the message consist of the five binary digits $(g_0, g_1, g_2, g_3, g_4)$. The code space $\mathfrak{S}_1^4$ may be regarded as generated by the four vectors x_1, x_2, x_3, x_4 and the identity vector I which may be written explicitly as follows:

$$\begin{aligned} x_1 &= (0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1) \quad , \\ x_2 &= (0\,0\,1\,1\,0\,0\,1\,1\,0\,0\,1\,1\,0\,0\,1\,1) \quad , \\ x_3 &= (0\,0\,0\,0\,1\,1\,1\,1\,0\,0\,0\,0\,1\,1\,1\,1) \quad , \\ x_4 &= (0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,1\,1\,1\,1\,1\,1) \quad , \\ I &= (1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1) \quad . \end{aligned} \tag{9}$$

The 32 vector codes of $\mathfrak{S}_1^4$ can be obtained by scalar multiplication of the vectors of (9) by the message digits g_0, g_1, g_2, g_3, g_4 in accordance with (8). For example, the message (0 1 1 0 0) has the code vector $g_1x_1 \oplus g_2x_2$ or

$$(0\,1\,1\,0\,0\,1\,1\,0\,0\,1\,1\,0\,0\,1\,1\,0) \quad .$$

Each of the 32 codes will be a distance of at least eight from each other.

In order to practically generate the above code, one should note that the vector x_1 is the sequence of digits generated by the least significant binary stage B_1 of a binary counter of scale sixteen; x_2 is obtained from the second stage B_2; x_3 from the third stage B_3; and x_4 from the final stage B_4, as the counter goes through one period of its operation. If the message $(g_0, g_1, g_2, g_3, g_4)$ is stored in a binary register with stages A_0, A_1, A_2, A_3, A_4, then the switching function

$$C = A_0 \oplus A_1B_1 \oplus A_2B_2 \oplus A_3B_3 \oplus A_4B_4$$

will generate the code sequentially during one period of operation of the binary counter.

If one of the above codes of $\mathfrak{S}_1^4$ is corrupted during transmission so that no more than three errors are made, it is evidently possible by the previous discussion of this section to somehow extract the original message from the corrupted received code. The method by which this extraction may be accomplished will be shown by example in the next section and in general in the last section. It should be clear from the above example how the vectors of $\mathfrak{S}_r^m$ may be generated for arbitray r and m where $r \leq m$.

IV. Decoding Corrupted Codes of $\mathfrak{S}_r^m$ by a Majority Testing of Redundancy Relations

Let us first consider the coding space $\mathfrak{S}_1^3$. By (7), the vector of this space has the form

$$g_0 I \oplus g_1x_1 \oplus g_2x_2 \oplus g_3x_3 \quad . \tag{10}$$

The message will consist of the four binary digits (g_0, g_1, g_2, g_3), and the generating vectors of the space are

$$\begin{aligned} x_1 &= (0\,1\,0\,1\,0\,1\,0\,1) \quad , \\ x_2 &= (0\,0\,1\,1\,0\,0\,1\,1) \quad , \\ x_3 &= (0\,0\,0\,0\,1\,1\,1\,1) \quad , \\ I &= (1\,1\,1\,1\,1\,1\,1\,1) \quad . \end{aligned} \tag{11}$$

By (6) we have the following set of relations for the message digits g_j in terms of f_k, the code digits.

$$g_0 = f(0, \ldots 0) = f_0 \quad , \qquad \Delta_{12} f(0\ldots) = f_0 \oplus f_1 \oplus f_2 \oplus f_3 = 0 \quad ,$$

$$g_1 = \Delta_1 f(0\ldots) = f_0 \oplus f_1 \quad , \qquad \Delta_{13} f(0\ldots) = f_0 \oplus f_1 \oplus f_4 \oplus f_5 = 0 \quad ,$$

$$g_2 = \Delta_2 f(0\ldots) = f_0 \oplus f_2 \quad , \qquad \Delta_{23} f(0\ldots) = f_0 \oplus f_2 \oplus f_4 \oplus f_6 = 0 \quad ,$$

$$g_3 = \Delta_3 f(0\ldots) = f_0 \oplus f_4 \quad , \qquad \Delta_{123} f(0\ldots) = \bigoplus_{i=0}^{7} f_i = 0 \quad . \tag{12}$$

By (12) there are four relations which g_1 satisfies,

$$g_1 = f_0 \oplus f_1 = f_2 \oplus f_3 = f_4 \oplus f_5 = f_2 \oplus f_3 \oplus f_4 \oplus f_5 \oplus f_6 \oplus f_7 \quad .$$

By substituting the second and third relations into the fourth relation, we have

$$g_1 = g_1 \oplus g_1 \oplus f_6 \oplus f_7 = 0 \oplus f_6 \oplus f_7 = f_6 \oplus f_7 \quad .$$

Thus we obtain the four independent and disjoint relations for g_1 ,

$$g_1 = f_0 \oplus f_1 = f_2 \oplus f_3 = f_4 \oplus f_5 = f_6 \oplus f_7 \quad .$$

These four relations are disjoint in the sense that no two of the relations have variables in common. In a similar manner, we may obtain four independent and disjoint relations for both g_2 and g_3 so that g_1, g_2, g_3 may be expressed as

$$g_1 = f_0 \oplus f_1 = f_2 \oplus f_3 = f_4 \oplus f_5 = f_6 \oplus f_7 \quad ,$$

$$g_2 = f_0 \oplus f_2 = f_1 \oplus f_3 = f_4 \oplus f_6 = f_5 \oplus f_7 \quad ,$$

$$g_3 = f_0 \oplus f_4 = f_1 \oplus f_5 = f_2 \oplus f_6 = f_3 \oplus f_7 \quad .$$

Let us now suppose that the received code is the vector $(f_0, f_1, \ldots f_7)$. If there were no error in transmission of the code, all of the above relations would hold. If there were one error, three out of four of the relations would hold. If there were two errors, at least two of the g_j's would have two out of four incorrect relations. Then g_1, g_2, g_3 may be determined uniquely if one or no error occurred during transmission, and two errors may always be detected by making a majority test on the arithmetic sum of the values of the four relations for each $g_j (j = 1,2,3)$. In order to state this criterion more explicitly, let the values of the four relations for g_j be denoted by $r_{j1}, r_{j2}, r_{j3}, r_{j4}$ for $(j = 1,2,3)$, and let S_j be the arithmetic sum of $r_{j1}, r_{j2}, r_{j3}, r_{j4}$ or

$$S_j = \sum_{i=1}^{4} r_{ji} \quad .$$

Then the majority decision test for g_j is

$$\begin{aligned} &g_j = 0 && \text{if } 0 \le S_j < 2 \quad , \\ &g_j \text{ is indeterminate} && \text{if } S_j = 2 \quad , \\ &g_j = 1 && \text{if } 2 < S_j \le 4 \text{ for } (j = 1,2,3) \quad . \end{aligned} \tag{13}$$

With the assumption that the received code is no more than two digits in error, the majority test (13) will determine g_1,g_2,g_3 uniquely for only one or no errors, and reject the code as meaningless in the case of two errors. In the case of one error or less, g_1,g_2,g_3 may be assumed now to be determined; it remains to determine g_0. In order to find g_0, note that if, as g_1,g_2,g_3 are found, the vectors g_1x_1,g_2x_2,g_3x_3 are added successively to the received vector, by (10) we will end with either the vector g_0I in the case of no error or with a vector of distance one from g_0I. Thus to detect g_0 the following majority decision test will suffice:

$$g_0 = 0 \quad \text{if} \quad \sum_{i=0}^{7} m_i < 4 \quad ,$$

$$= 1 \quad \text{if} \quad \sum_{i=0}^{7} m_i > 4 \quad , \tag{14}$$

where m_i are the digits of the code after extraction of digits g_1,g_2,g_3 in accordance with the above procedure.

The above method of decoding may be illustrated by the following example: Suppose that the message sent was (1 0 1 1), and that during transmission an error was made in the fifth digit of the original code (1 1 0 0 0 0 1 1) so that the received code had the form (1 1 0 0 1 0 1 1). We first test for g_1,g_2,g_3 by (12) and find $g_1 = 0$, $g_2 = 1$ and $g_3 = 1$. Using (11), we add $g_1x_1 \oplus g_2x_2 \oplus g_3x_3$ to the code, obtaining

$$0(0\,1\,0\,1\,0\,1\,0\,1) \oplus (0\,0\,1\,1\,0\,0\,1\,1) \oplus (0\,0\,0\,0\,1\,1\,1\,1) \oplus (1\,1\,0\,0\,1\,0\,1\,1)$$

$$= (1\,1\,1\,1\,0\,1\,1\,1) = (m_0,m_1,m_2,\dots m_3) \quad .$$

Finally, by (14)

$$g_0 = 1 \quad , \quad \text{since} \quad \sum_{i=0}^{7} m_i = 7 > 4 \quad .$$

Although $\mathfrak{P}_1^3$ is none other than an example of a set of one-error-correcting and two-error-detecting codes of the type described by Hamming in Ref. 1, the method of decoding considered above is different. Our procedure of decoding is advantageous in that it may be generalized in a natural way to include any of the coding spaces $\mathfrak{P}_r^m$ of the second section of this paper. Before we consider the generalization by further examples, let us note a tabular way of representing the redundancy relations.

If the digits or variables of each relation are connected by lines for each of the vectors x_1,x_2,x_3 as

$$x_1 = (0\ 1\ 0\ 1\ 0\ 1\ 0\ 1) \quad ,$$

$$x_2 = (0\ 0\ 1\ 1\ 0\ 0\ 1\ 1) \quad ,$$

$$x_3 = (0\ 0\ 0\ 0\ 1\ 1\ 1\ 1) \quad , \tag{15}$$

the relations of (12) become almost self-evident by their simplicity with respect to order and symmetry. This simplicity makes it possible to discover the redundancy relations for more general spaces $\mathfrak{P}_r^m$ without resorting to the algebraic approach used above.

As a second example of our decoding procedure, consider the coding space $\mathfrak{P}_1^4$ introduced in the latter part of the preceding section. Each vector of this space has the form of (8), where the generating vectors are x_1,x_2,x_3,x_4 and I of (9). The first-degree redundancy relations may be determined in a manner similar to the above example and represented in a tabular manner similar to (15) as follows:

$$x_1 = (0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1)\ ,$$

$$x_2 = (0\ 0\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 1)\ ,$$

$$x_3 = (0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1)\ ,$$

$$x_4 = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1)\ .$$

(16)

For instance, the eight independent and disjoint relations for g_1 are

$$g_1 = f_{2i} \oplus f_{2i+1} \quad \text{for } (i = 0,1,\ldots 7) \quad .$$

If the eight values of the redundancy relations for g_j are labeled $r_{j1}, r_{j2}, \ldots r_{j8}$ for $(j = 1,2,3,4)$, and S_j is defined by

$$S_j = \sum_{i=1}^{8} r_{ji} \quad ,$$

then, by an argument similar to that used in the previous example, the majority decision test for g_j is as follows:

$$\begin{aligned} g_j &= 0 && \text{if } 0 \le S_j < 4 \quad , \\ g_j &\text{ is indeterminate} && \text{if } S_j = 4 \quad , \\ g_j &= 1 && \text{if } 4 < S_j \le 8 \quad \text{for } (j = 1,2,3,4). \end{aligned} \tag{17}$$

In order to determine g_0, we first add the determined vectors $g_j x_j$ to the received message, assuming, of course, that no g_j is indeterminate, and we are left with the zero-degree polynomial Φ_0^4, possibly corrupted by errors. If there had been no errors, there would be sixteen zero-degree relations which g_0 satisfies, or

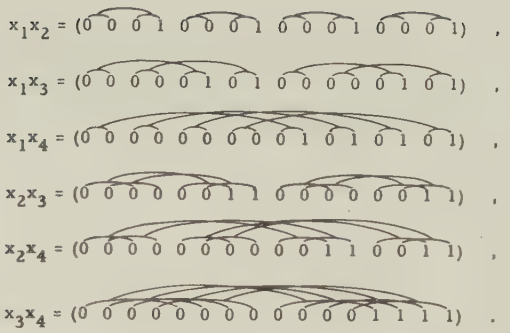
$$g_0 = m_j \quad \text{for } (j = 0,1,2,\ldots 15) \quad ,$$

where, as in (14), m_j are the digits of the code after extraction of g_1, g_2, g_3 and g_4. Thus g_0 is determined by the majority decision test

$$\begin{aligned} g_0 = 0 \quad &\text{if} \quad \sum_{i=0}^{15} m_i < 8 \quad , \\ = 1 \quad &\text{if} \quad \sum_{i=0}^{15} m_i > 8 \quad . \end{aligned} \tag{18}$$

For the above example three errors may be made in the code and the correct message obtains. If four errors are made, some of the message digits are indeterminate. It is of some interest to note that, for some cases of five errors in the code, the message may be extracted correctly. For example, suppose that the message was (0 0 0 0 0) and that the received code was (1 1 0 0 1 0 1 0 1 0 0 0 0 0 0 0). Clearly, the correct message will be extracted from this code by the above procedure.

As a final example of coding and decoding scheme, consider Φ_2^4. This space is generated by x_1, x_2, x_3, x_4 of (16) and I, as well as the quadratic variables $x_1x_2, x_1x_3, x_1x_4, x_2x_3, x_2x_4, x_3x_4$. The latter six vectors may be presented in the following tabular manner.



(19)

The messages for this example will be 11 binary digit numbers of the form $(g_0, g_1, g_2, g_3, g_4, g_{12}, g_{13}, g_{14}, g_{23}, g_{24}, g_{34})$. Each code will be sent as a vector of the form

$$g_0 \oplus g_1x_1 \oplus g_2x_2 \oplus g_3x_3 \oplus g_4x_4 \oplus g_{12}x_1x_2 \oplus g_{13}x_1x_3 \oplus g_{14}x_1x_4$$

$$\oplus g_{23}x_2x_3 \oplus g_{24}x_2x_4 \oplus g_{34}x_3x_4 \quad .$$

The second-degree coefficients g_{ij} of the received message are extracted first with a majority decision based on the redundancy relations illustrated in (19). Next, assuming that no indeterminacy occurred in the second-degree coefficients, the vectors $g_{ij}x_ix_j$ are added to the received code, after which we are left with a residual code from which the first-degree coefficient g_0 may be determined by test (18) after adding the vectors $g_1x_1, g_2x_2, g_3x_3, g_4x_4$ to the residual code.

This example illustrates the general principle of decoding the particular class of codes under consideration. The highest degree coefficients of a received code are extracted first; then these terms of the polynomial are subtracted out of the code, thereby leaving a residual code of the next lower degree than the original code in the special case of no errors. The operation is repeated over and over on the successive residual codes until either an indeterminacy occurs or until g_0 is extracted.

The relations of (19) illustrate the fact that there are four redundancy relations each of four variables for the second-degree coefficients g_{ij}. For example, the redundancy relations for g_{12} are

$$g_{12} = f_{4i} \oplus f_{4i+1} \oplus f_{4i+2} \oplus f_{4i+3} \quad \text{for } (i = 0, 1, 2, 3) \quad . \tag{20}$$

In general, these relations will allow only one error; two errors will lead to indeterminacy. This is another example of Hamming's one-error-correction and two-error-detection codes.

It should be noted that the majority decision tests used in the above examples were, in general, overdeterminate. For instance, in the first example, if one error had been made, no more than one error would remain in the residual code after determining g_1, g_2, g_3. On the other hand, if two errors had occurred, the process of extraction would have ended before g_0 could be determined. Thus a test of only the following type would be necessary:

$$g_0 = 0 \quad \text{if } m_{i1} + m_{i2} + m_{i3} \leq 1 \quad ,$$

$$g_1 = 1 \quad \text{if } m_{i1} + m_{i2} + m_{i3} \geq 2 \quad ,$$

where i_1, i_2, i_3 are any three distinct numbers between zero and seven, inclusive. Refinements such as this, however, do not destroy the validity of the previous tests.

V. THE GENERAL DECODING PRINCIPLE

To study the general decoding scheme, illustrated by example in Section IV, it will be necessary to consider the general multinomial expansion formula (6) more carefully. Let us first define the multiple differences, used in (6) in more detail.

As in (6), $f(i_1, \ldots i_m)$ is defined as

$$f(i_1,\ldots i_m) = f_j \quad \text{when } j = \sum_{k=1}^{m} i_k 2^{k-1} \quad \text{for } (i_k = 0, 1) \quad . \tag{21}$$

The general multiple partial difference

$$\overset{p}{\underset{k_1,k_2,\ldots k_p}{\Delta}} f(i_1,i_2,\ldots i_m)$$

is defined inductively as

$$\overset{}{\underset{k}{\Delta}} f(i_1,\ldots i_m) = f(i_1,\ldots i_{k-1},i_k \oplus 1, i_{k+1},\ldots i_m) \oplus f(i_1,\ldots i_k,\ldots i_m)$$

$$\overset{p}{\underset{k_1k_2,\ldots k_p}{\Delta}} f(i_1,\ldots i_m) = \overset{p-1}{\underset{k_1,\ldots k_{p-1}}{\Delta}} f(i_1,\ldots i_{k_{p-1}},i_{k_p} \oplus 1, i_{k_{p+1}},\ldots i_m)$$

$$\oplus \overset{p-1}{\underset{k_1,\ldots k_{p-1}}{\Delta}} f(i_1,\ldots i_m) \tag{22}$$

With these definitions it is possible to prove by induction the validity and uniqueness of expansion (6) for any Boolean algebra of m variables, and in particular, for the Boolean algebra of 2^m dimensional vectors as described in Section II.

One evident consequence of (21) is the identity

$$f(i_1,\ldots i_{k-1}i_k \oplus 1, i_{k+1},\ldots i_m) = f_{i+(-1)^{i_k}2^{k-1}} \quad . \tag{23}$$

By the use of (23) it is possible to write (22) explicitly in terms of the f_i as

$$\overset{}{\underset{k}{\Delta}} f(i_1,\ldots i_m) = f_i \oplus f_{i+(-1)^{i_{k2}}2^{k-1}}$$

and

$$\overset{p}{\underset{k_1,k_2,\ldots k_p}{\Delta}} f(i_1,\ldots i_m) = \sum_{i=1}^{2^{p-1}} f_{j_i} \oplus \sum_{i=1}^{2^{p-1}} f_{j_i+(-1)^{i_{k_p2}}2^{k_p-1}}$$

where

$$\overset{p-1}{\underset{k_1,k_2,\ldots k_{p-1}}{\Delta}} f(i_1,\ldots i_m) = \sum_{i=1}^{2^{p-1}} f_{j_i} \quad \text{and} \quad j_i \neq j_s+(-1)^{i_{k_p}2^{k_p-1}}$$

$$\text{for } (i,s = 1,\ldots 2^{p-1}) \quad . \tag{24}$$

We are now in a position to prove the following fundamental theorem on which the general decoding principle of the class of codes under consideration rests.

Theorem B:- Each highest or r-th degree coefficient of any vector or polynomial f of Φ_r^m satisfies exactly 2^{m-r} disjoint relations where each relation has precisely the form

$$\sum_{k=1}^{2^r} f_{i_k} \quad ,$$

where i_k are distinct numbers from the set $(0,1,2,\ldots 2^m - 1)$ for $(k = 1,2,\ldots 2^r)$. Disjointness of relations means that no two relations have variables f_i in common.

Proof:- Choose m and r. By (6),(7) and (24), the highest degree coefficients for an f of Φ_r^m are

$$g_{k_1\ldots k_r} = \Delta^r_{k_1k_2\ldots k_r} f(0,\ldots 0) = \sum_{i=1}^{2^r} f_{j_i} \quad , \tag{25}$$

where k_j are distinct integers from the set $(1,2,\ldots m)$ for $(j = 1,\ldots r)$, and j_i are distinct integers from the set $(0,1,\ldots 2^m - 1)$ for $(i = 1,2,\ldots 2^r)$. Moreover,

$$\Delta_{k_1\ldots k_r n_1 n_2\ldots n_t} f(0,\ldots 0) = 0 \tag{26}$$

for $t \geq 1$, and k_j and n_l are distinct integers from the set $(1,2,\ldots m)$ for $(j = 1,\ldots t)$.

Let $k_1,k_2\ldots k_r$ be a distinct set of integers from the set $(1,2,\ldots m)$. Then by (26) and (22),

$$\Delta^{r+1}_{k_1\ldots k_r n_1} f(0,\ldots 0) = \Delta^r_{k_1\ldots k_r} f(0,\ldots 0) \oplus \Delta^r_{k_1\ldots k_r} f(0,\ldots 1,\ldots 0) = 0 \tag{27}$$

where n_1 is any one of the m-r integers from the set $(1,2,\ldots m)$ which is distinct from the integers $(k_1,k_2,\ldots k_r)$. Thus, by (24) and (25), we have exhibited m-r new relations of the form required by the theorem. Each of these new relations is distinguished by the fact that the digit one appears only in the n_1-th position of the function $f(i_1,\ldots i_m)$ operated on by

$$\Delta^r_{k_1\ldots k_r}$$

Now define $f\left[n_1,n_2,\ldots n_t\right]$ to be $f(i_1,i_2,\ldots i_m)$ with $i_k = 1$ for $k = n_1,n_2,\ldots n_t$ and $i_k = 0$ otherwise. The theorem will be proved by induction on the subscript of n. Assume therefore that

$$\Delta^{r+s-1}_{k_1k_2\ldots k_r n_1 n_2\ldots n_{s-1}} f(0,0,\ldots 0) = \Delta^r_{k_1\ldots k_r} f(0,0,\ldots 0) \oplus \Delta^r_{k_1\ldots k_r} f\left[n_1,n_2\ldots n_{s-1}\right] . \tag{28}$$

Now, by (22) and (26) and the induction hypothesis (28),

$$\overset{r+s}{\underset{k_1\ldots k_r n_1\ldots n_s}{\Delta}} f(0,0,\ldots 0) = \underset{n_s}{\Delta}\left(\overset{r+s-1}{\underset{k_1\ldots k_r n_1\ldots n_{s-1}}{\Delta}} f(0,0,\ldots 0)\right)$$

$$= \underset{n_s}{\Delta}\left(\overset{r}{\underset{k_1\ldots k_r}{\Delta}} f(0,\ldots 0) \oplus \overset{r}{\underset{k_1\ldots k_r}{\Delta}} f\left[n_1,\ldots n_{s-1}\right]\right)$$

$$= \overset{r}{\underset{k_1,\ldots k_r}{\Delta}} f(0,0,\ldots 0) \oplus \overset{r}{\underset{k_1\ldots k_r}{\Delta}} f\left[n_1,\ldots n_{s-1}\right]$$

$$\oplus \overset{r}{\underset{k_1\ldots k_r}{\Delta}} f\left[n_s\right] \oplus \overset{r}{\underset{k_1\ldots k_r}{\Delta}} f\left[n_1,\ldots n_s\right] = 0 \quad .$$

Now, by (27) and (28), the two middle terms are equal to

$$\overset{r}{\underset{k_1\ldots k_r}{\Delta}} f(0,0,\ldots 0) \quad ,$$

and therefore their sum modulo 2 is zero. Hence

$$\overset{r+s}{\underset{k_1\ldots n_s}{\Delta}} f(0,\ldots 0) = \overset{r}{\underset{k_1\ldots k_r}{\Delta}} f(0,\ldots 0) \oplus \overset{r}{\underset{k_1\ldots k_r}{\Delta}} f\left[n_1,\ldots n_s\right] = 0 \quad ,$$

and the induction is complete. The theorem is proved when we observe that the relation

$$\overset{r+s}{\underset{k_1\ldots n_s}{\Delta}} f(0,\ldots 0) = 0 \text{ contributes } \binom{m-r}{s} \text{ distinct relations,}$$

$$\underset{k_1\ldots k_r}{\Delta} f(0,\ldots 0) = \underset{k_1\ldots k_r}{\Delta} f\left[n_1,n_2\ldots n_s\right] \quad ,$$

since there are $\binom{m-r}{s}$ ways of choosing s integers from m-r integers. Using all the relations (26) for the particular set $k_1\ldots k_r$ and t = 1 to t = m-r and the relation (25), we get

$$1 + \sum_{t=1}^{m-r} \binom{m-r}{t} = 2^{m-r}$$

distinct relations for $g_{k_1,k_2,\ldots k_r}$. Since these relations exhaust all variables f_{i_k}, the theorem is proved.

The above theorem shows that the generalization of the decoding principle, discussed in the last section obtains. The majority decision test for the general case can clearly be used to extract the r-th degree coefficients of Φ_r^m, where the relations used for the test are the 2^{m-r} relations of Theorem B. The (r-1)-th degree coefficients are then extracted the same way after the determined r-th order terms have been subtracted or added into the received code. This process is continued for the r-2,r-3,... degree coefficients until the message is extracted or an indeterminacy is reached.

VI. Concluding Remarks

Since there are $\binom{m}{j}$ j-th degree coefficients $g_{i_1 i_2 \ldots i_j}$ in expansion (5), there must be

$$N = \sum_{i=0}^{r} \binom{m}{i}$$

coefficients in each polynomial (7) of the coding space Φ_r^m. The coefficients of (7) constitute the message sent, thus each code of Φ_r^m contains N bits of message information. Since each element of Φ_r^m is a vector of dimension 2^m, there are 2^m-N bits of the code used to supply redundancy.

In order to illustrate the relationship of the number of message bits to number of errors corrected, consider the coding space Φ_4^7. By (29) each code of (29) has 99 bits of message information for a code of 128 bits. By Section III at least

$$2^{m-r-1} - 1 = 2^{7-4-1} - 1 = 3$$

bits of error in the code can be corrected. By Section IV and Section V four bits of error will lead undoubtedly to an indeterminacy in the message and it is likely that in some cases of five errors the correct message will be extracted by the majority decision process. Further examples of the numerical relationship of message bits to number of errors corrected may be constructed in a similar manner.

Attempts have been made with little success to investigate the structure of the complete convex set S of points, containing an element σ of Φ_r^m, whose points correspond to the element σ under the majority decision test procedure of Section V. As the second example of Section IV shows, there are in general more points in S than in a Hamming sphere of radius 2^{m-r-1} containing σ. These attempts were motivated by a desire to show that the coding system discussed here would satisfy Shannon's fundamental theorem for a discrete channel with noise (Theorem 11 in Ref. 3). So far, this fact has not been shown.

There are two generalizations of the codes discussed in this paper. In Ref. 2 Muller discusses generalizations of the binary codes, discussed here, for lengths other than 2^m. Another generalization is possible where the polynomials considered here are considered over a field of characteristic other than two; i.e., ternary codes, etc. It will not be the purpose of this paper to investigate these generalizations.

ACKNOWLEDGMENTS

The author expresses his appreciation to E. B. Rawson for his assistance in the construction of the second example of Section 4; to G. P. Dinneen for his help in the simplification of Theorem B; and to T. A. Kalin, W. B. Davenport, D. E. Muller, and O. G. Selfridge for several useful discussions.

REFERENCES

1. R. W. Hamming, Bell System Tech. J. 26, No. 2, 147 (April 1950).
2. D. E. Muller, "Metric Properties of Boolean Algebra and Their Application to Switching Circuits," Report No. 46, Digital Computer Laboratory, Univ. of Illinois (April 1953).
3. D. E. Shannon, "A Mathematical Theory of Communication," Bell System Tech. J. 27, (July,October 1948).

CODING FOR CONSTANT-DATA-RATE SYSTEMS*†

Richard A. Silverman and Martin Balser
Lincoln Laboratory, M.I.T.
Lexington, Mass.

A. INTRODUCTION

We consider a communication system in which data consisting of sequences (known as words) of binary digits are transmitted at a predetermined constant rate. (For our purposes, a binary digit is one of two electrical signals of duration T and bandwidth W.) The nature of the data and the manner in which they are translated into words are irrelevant to this discussion. For example, the data may be English letters, numbers, etc., reduced to sequences of five binary digits each for use in a teletype system, or they may be conventional symbols representing entire messages.

A basic problem of coding is to reduce the average rate of incorrectly received words as much as possible. Accordingly, additional digits are added to the word for the purposes of error detection or correction. The assumption that the words are sent at a constant rate requires that each binary digit be shortened by such an amount that the coded words (message digits plus check digits) have the same duration as the original uncoded words. This shortening of each digit increases its probability of error and, consequently, the probability of error per word. On the other hand, the coding imposes constraints on the digits composing a word, so that errors may show up as inconsistencies and may in many cases be corrected. This tends to reduce the probability of error per word. The efficacy of a code depends on how much the second effect outweighs the first.

The simplest code of all consists of an extra digit selected to make the sum of all the digits in the coded word even (or odd). If the sum of the digits of the received word has the wrong parity, an odd number of errors is known to have been made. There is, however, no indication of the correct replacement for the mistaken word, unless some dependence between separate words (such as the redundancy of printed English[1]) is exploited.

Hamming[2] has devised a code that corrects all single errors. It consists of adding k suitably chosen check digits to the m message digits. If another digit is added, double errors can be detected as well as single errors corrected.[2] In this paper, we describe a new single-error-correcting code (the Wagner code) and evaluate its performance in a constant-data-rate system, particularly as compared with that of the Hamming code.

The principle of the Wagner code is readily extended to the construction of multiple-error-correcting codes. We evaluate the performance in a constant-data-rate system of two such codes and of a code recently developed by I.S. Reed.

B. DESCRIPTION OF THE WAGNER CODE

In this study, we are concerned with communication systems that transmit words consisting of binary digits. A binary digit is one of two electrical signals $x_1(t)$ and $x_2(t)$ of duration T and bandwidth W. Let $p(x_i/y)$ be the (a posteriori) probability that if y is received, x_i was

*This paper is a condensation of two papers of this title, Part I, A New Error-Correcting Code, by R.A. Silverman and M. Balser, and Part II, Multiple-Error-Correcting Codes, by M. Balser and R.A. Silverman, which will appear in the Proceedings of the I.R.E. Further details and derivations which are omitted from this condensation are to be found in the more complete papers.

†The research in this document was supported jointly by the Army, Navy and Air Force under contract with the Massachusetts Institute of Technology.

sent, and let Δp be $p(x_1/y) - p(x_2/y)$. In the absence of any constraints on the digits composing a word or of dependence between the words themselves, the receiver can compute only $p(x_1/y)$ and $p(x_2/y)$, and for each digit choose x_1 or x_2, depending on whether $p(x_1/y)$ or $p(x_2/y)$ is the larger (or, equivalently, whether Δp is positive or negative). The error-correcting code that we shall describe (named the Wagner code after C. A. Wagner of this laboratory, who suggested the basic idea) enables us to use some of the information presented by the magnitudes of the Δp's* by introducing a constraint on the digits composing a word. This information is ignored by more conventional codes.

In the Wagner code, a transmitted word consists of a sequence of m message digits and an additional digit used as a parity check. As each of the perturbed digits y arrives at the receiver, the a posteriori probabilities $p(x_1/y)$ and $p(x_2/y)$ are calculated. Each digit of the received sequence is tentatively identified as x_1 or x_2, depending on whether $p(x_1/y)$ or $p(x_2/y)$ is the larger, and the values of the a posteriori probabilities are stored in a memory for the duration of a word. The sequence thus obtained is checked for parity. If the parity is correct, the word is printed as received. If the parity check fails, the digit for which the difference Δp between a posteriori probabilities is the smallest is considered the digit most in doubt, and the word is printed with this digit altered. The receiver then clears the stored values of the probability differences from the memory and proceeds to the next word.

Thus we may characterize the Wagner code as one which probably corrects single errors. (Multiple errors are always printed incorrectly.) However, as we shall see, it can be more effective in a constant-data-rate system than a code that corrects all single errors (such as the Hamming code).

The a posteriori probabilities $p(x_1/y)$ and $p(x_2/y)$ are functions of the random received waveform y(t) and therefore are themselves random variables. The calculation of their distributions is, in general, very difficult. For simplicity, we shall consider the case where the two transmitted signals have equal energy and equal a priori probabilities and are perturbed by the addition of white Gaussian noise. It has been shown[4] that for this case

$$p(x_1/y) = \beta \exp\left[\gamma \int_0^T x_1(t)\, y(t)\, dt\right] \tag{1}$$

and

$$p(x_2/y) = \beta \exp\left[\gamma \int_0^T x_2(t)\, y(t)\, dt\right] \quad ,$$

where β and γ are constants. Thus the transmitted signal with the larger correlation has the larger a posteriori probability. Equivalently, we may write

$$\frac{p(x_1/y)}{p(x_2/y)} = \exp\left[\gamma(z_1 - z_2)\right] \quad , \tag{2}$$

where

$$z_1 = \int_0^T x_1(t)\, y(t)\, dt \qquad \text{and} \qquad z_2 = \int_0^T x_2(t)\, y(t)\, dt \tag{3}$$

* The signs of the Δp's are used in making the tentative identification of the transmitted word.

From Eq. (2), we see that the smaller the difference $\Delta z = z_1 - z_2$ between the correlation integrals, the closer to unity the ratio of a posteriori probabilities and, consequently, the smaller the difference Δp between the two probabilities. Thus, if the parity check fails, the digit that should be changed (as the one most in doubt) is the one for which Δz is the smallest.

It can be shown that z_1 and z_2 are normally distributed random variables (with means c_1 and c_2, and variances σ_1 and σ_2), so that calculations are especially simple for the correlation detector.* Moreover, under the assumptions made above, the correlation detector is equivalent to the probability detector. Therefore, it is assumed in what follows that detection is by correlation.

C. ANALYSIS OF THE WAGNER CODE

1. Probability of Error Per Digit

As noted above, the correlation integrals z_1 and z_2 (corresponding to the signal that was sent and the signal that was not sent, respectively) are random variables with probability densities

$$W(z_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{(z_1 - c_1)^2}{2\sigma_1^2}\right] , \tag{4}$$

and

$$W(z_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{(z_2 - c_2)^2}{2\sigma_2^2}\right] . \tag{5}$$

If x_1 and x_2 are suitably chosen, z_1 and z_2 may be regarded as statistically independent random variables. Accordingly, the probability density of finding a separation $\Delta z = z_1 - z_2$ is

$$W(\Delta z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\Delta z - \Delta c)^2}{2\sigma^2}\right] , \tag{6}$$

where $\Delta c = c_1 - c_2 > 0$ and $\sigma^2 = \sigma_1^2 + \sigma_2^2$. If Δz is negative, then selecting as the transmitted signal the signal giving the larger correlation integral will result in an error. Thus the probability of error per digit is

$$p(a) = \int_{-\infty}^{0} W(\Delta z)\, d\Delta z = \frac{1}{2}(1 - \operatorname{erf} a) = 1 - q(a) , \tag{7}$$

where $a = \Delta c/\sqrt{2}\sigma$. The parameter a, which is proportional to the signal-to-noise ratio of the

*For simplicity of notation, we shall always use the subscript "one" for the signal that was transmitted. Thus $c_1 > c_2$, and $z_1 < z_2$ results in an error.

correlator difference, is the significant parameter in the calculations that follow.

Since we do not know which digit was actually sent, we do not know the sign of Δz. From Eq.(6), the joint probability that the correlator difference lies between $|\Delta z|$ and $|\Delta z| + d|\Delta z|$, and that the larger correlation integral corresponds to the transmitted signal, is $d|\Delta z|$ times

$$W(|\Delta z|,\text{right}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(|\Delta z| - \Delta c)^2}{2\sigma^2}\right] \quad . \tag{8}$$

On the other hand, the joint probability that the correlator difference lies between $|\Delta z|$ and $|\Delta z| + d|\Delta z|$, and that the larger correlation integral does not correspond to the transmitted signal, is $d|\Delta z|$ times

$$W(|\Delta z|,\text{wrong}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(|\Delta z| + \Delta c)^2}{2\sigma^2}\right] \quad . \tag{9}$$

2. Probability of Correcting a Single Error

Suppose that the received word has n digits. Since the parity check fails if a single error is made, and since then the Wagner code changes the digit with the smallest $|\Delta z|$, the probability $\Pi_n(a)$ that a single error is made and is corrected by the Wagner code is just the probability that the digit with the smallest $|\Delta z|$ is incorrect and that the $n - 1$ other digits are correct. $\Pi_n(a)$ can be calculated as follows.

Let $|\Delta z_i|$ be the correlator difference for the i-th digit. Since the $|\Delta z_i|$ are independent random variables, the joint probability that the first digit, with correlator difference between $|\Delta z_1|$ and $|\Delta z_1| + d|\Delta z_1|$, is wrong, and that all the other digits, with correlator differences between $|\Delta z_2|$ and $|\Delta z_2| + d|\Delta z_2|$, ..., $|\Delta z_n|$ and $|\Delta z_n| + d|\Delta z_n|$, are right, is

$$W(|\Delta z_1|,\text{wrong}) \prod_{i=2}^{n} W(|\Delta z_i|,\text{right}) \prod_{i=1}^{n} d|\Delta z_i| \quad . \tag{10}$$

Thus the joint probability that the first digit is wrong, that the $n - 1$ other digits are correct, and that $|\Delta z_1| < |\Delta z_2| < \ldots < |\Delta z_n|$ is

$$\begin{aligned} &\int_0^\infty W(|\Delta z_n|,\text{right})\, d|\Delta z_n| \int_0^{|\Delta z_n|} W(|\Delta z_{n-1}|,\text{right})\, d|\Delta z_{n-1}| \\ &\ldots \int_0^{|\Delta z_3|} W(|\Delta z_2|,\text{right})\, d|\Delta z_2| \int_0^{|\Delta z_2|} W(|\Delta z_1|,\text{wrong})\, d|\Delta z_1| \quad . \end{aligned} \tag{11}$$

Since there are in all n! orderings of the n correlator differences, the joint probability $\Pi_n(a)$ that a single error is made in a word of n digits and that it is corrected by the Wagner code is given by

$$\begin{aligned} \Pi_n(a) = \frac{n!}{(\sqrt{\pi})^n} &\int_0^\infty \exp\left[-(x_n - a)^2\right] dx_n \int_0^{x_n} \exp\left[-(x_{n-1} - a)^2\right] dx_{n-1} \\ &\ldots \int_0^{x_3} \exp\left[-(x_2 - a)^2\right] dx_2 \int_0^{x_2} \exp\left[-(x_1 + a)^2\right] dx_1 \quad , \end{aligned} \tag{12}$$

where $a = \Delta c/\sqrt{2}\sigma$ as in Eq. (7). Equation (12) can be reduced to the following form, more suitable for computations.

$$\Pi_n(a) = \frac{n}{2^n} \sum_{i=1}^{n} \binom{n-1}{i-1} (-1)^{i+1} I_i(a) \tag{13}$$

where

$$I_n(a) = \frac{2}{\sqrt{\pi}} \int_0^{\infty} [\text{erf}\,(x-a)]^{n-1} \exp\,[-(x+a)^2]\,dx \quad . \tag{14}$$

D. PROBABILITY OF ERROR FOR WAGNER-CODED WORDS – COMPARISON WITH UNCODED AND HAMMING-CODED WORDS

The probability of error per word for a Wagner-coded word containing m message digits (n = m + 1 digits in all) is

$$P_W = 1 - q^{m+1}(a) - \Pi_{m+1}(a) \quad , \tag{15}$$

that is, the probability of error is one minus the sum of the probability that the word is received correctly and the probability that a single error is made and then corrected. We wish to compare P_W with P_U, the probability of error per word if no code is used, and P_H, the probability of error per word if the Hamming single-error-correcting code is used. Since we are concerned with constant-data-rate systems, the duration of the transmitted signals must be altered if coded words (message digits plus error-correcting digits) are to have the same duration as differently coded or uncoded words. Changing the signal duration changes the variance of the correlator difference and consequently the value of the parameter a and the probability of error per digit [see Eq. (7)]. For large TW (the only case we consider), it can be shown[5] that $a(= \Delta c/\sqrt{2}\sigma)$ is proportional to $T^{\frac{1}{2}}$. Using this result, we find that, for the same value of a used in Eq (15),

$$P_U = 1 - q^m\left(\sqrt{\frac{m+1}{m}}\, a\right) \tag{16}$$

and

$$P_H = 1 - q^{m+k}\left(\sqrt{\frac{m+1}{m+k}}\, a\right) - (m+k)\, q^{m+k-1}\left(\sqrt{\frac{m+1}{m+k}}\, a\right) p\left(\sqrt{\frac{m+1}{m+k}}\, a\right) \quad , \tag{17}$$

where k is the number of check digits required by the Hamming code. P_U, P_H, and P_W have been computed for values of m from 4 to 8 and for selected values of a. The results are given in Table I.

There is only a certain range of values of the signal-to-noise ratio of the correlator difference for which it is worth the effort to implement either of the error-correcting codes. For high signal-to-noise ratio, very few errors are made, and additional equipment is generally not justifiable. On the other hand, for low signal-to-noise ratio, multiple errors become too frequent, and single-error-correcting codes are of little use. Thus single-error-correcting codes are of considerable value for values of a from about 1.0 to 3.0.

TABLE I

Probabilities of error per word for uncoded, Hamming-coded, and Wagner-coded words containing m message digits				
m	α	P_U	P_H	P_W
4	1.0	0.209	0.191	0.143
	1.5	0.0349	0.0248	0.0115
	2.0	0.00313	0.00145	0.00030
	3.0	42×10^{-7}	6×10^{-7}	$< 10^{-7}$
5	1.0	0.269	0.310	0.190
	1.5	0.0493	0.0513	0.0164
	2.0	0.00486	0.00375	0.00045
	3.0	84×10^{-7}	25×10^{-7}	$< 10^{-7}$
6	1.0	0.325	0.335	0.236
	1.5	0.0641	0.0530	0.0220
	2.0	0.00673	0.00346	0.00062
	3.0	138×10^{-7}	17×10^{-7}	$< 10^{-7}$
7	1.0	0.377	0.362	0.282
	1.5	0.0789	0.0552	0.0281
	2.0	0.00871	0.00330	0.00082
	3.0	201×10^{-7}	12×10^{-7}	$< 10^{-7}$
8	1.0	0.425	0.388	0.326
	1.5	0.0937	0.0580	0.0347
	2.0	0.01075	0.00322	0.00103
	3.0	272×10^{-7}	9×10^{-7}	$< 10^{-7}$

As shown in Table I, in this range of values the probability of error of Wagner-coded words is considerably less than that of Hamming-coded words. For increasing word length, the advantage of the Wagner code diminishes from two causes: (1) the ratio k/m decreases with increasing m so that the length of the digits in the Hamming-coded word approaches those of the Wagner-coded words, thus narrowing the gap between the corresponding signal-to-noise ratios; and (2) the conditional probability of correcting a single error decreases with increasing m. Nonetheless, even for m = 8 and α = 2, we may expect only 103 errors per 100,000 words using the Wagner code, as compared with 322 errors per 100,000 Hamming-coded words and 1,075 errors per 100,000 uncoded words.

E. IDENTIFICATION OF THE SMALLEST CORRELATOR DIFFERENCE

Until now we have assumed that the receiver can pick out the smallest correlator difference with infinite precision. Suppose, however, that the equipment used in implementing the Wagner code is such that the smaller of two correlator differences within $\epsilon \Delta c$ of each other cannot be identified with certainty. We have found that even for a rather crude receiver, which cannot distinguish correlator differences lying within 0.1 Δc of each other, the percentage change in the probability of error per word is at most about two per cent in the region of interest. Thus the advantage of the Wagner code over the Hamming code does not depend on great precision of the correlators or the memory.

F. THE HAMMING-WAGNER CODE

We now extend the principle of the Wagner code to a double-error-correcting code. The following procedure appears best as a first attempt. Further check digits are added to the Wagner-coded word; these reveal double as well as single errors. If a double error is detected, we change the two digits of the stored word with the smallest correlator differences. If a single error is detected, we change only the smallest correlator difference.

The success of this scheme requires a system of check digits which indicates both single and double errors, and further allows them to be distinguished. The geometrical model of message space is well suited for examining the possibility of setting up such check digits.* Referring to Fig.1, we see that if both single and double errors in possible transmitted points (such as P_1 and P_2) are to be detectable, and if single errors are to be distinguishable from double errors, every such pair of points must be separated by a distance of 4 or more. For then, a single error in P_1 sends it to a neighboring point like S_1, where it can be stated with certainty to have come either from P_1 by a change in one digit, or from some other possible transmitted message by a change in three or more digits. Similarly, a single error in P_2 sends it to a neighbor like S_2. On the other hand, a double error in either P_1 or P_2 may correspond to a received point like D, at a distance of 2 from both. Unless there are at least three points between all pairs of possible transmitted points, a double error in P_1 (say) is indistinguishable from a single error in P_2 (or some other transmitted point), so that we do not know whether to correct one or two digits in the received word.

Now in a Hamming single-error-correcting, double-error-detecting code,[2] all transmitted messages are separated by at least a distance of 4. This is just the separation required for successful operation of a Wagner code that corrects both single and double errors. Thus the number of check digits needed to correct all single errors before applying the Wagner procedure to double errors is the same as the number required to apply the Wagner procedure to both single and double errors. This suggests a "Hamming-Wagner" code, which is obviously better than the corresponding "Wagner-Wagner" code.

We thus arrive at a code that is like the Hamming single-error-correcting, double-error-detecting code, except that if the extra check digit indicates a double error, we change the two digits with the smallest correlator differences. The analysis of this Hamming-Wagner code is completely analogous to that of the simple Wagner code.

The probability of error per Hamming-Wagner-coded word is

$$P_{HW}(a) = 1 - q^{m+k+1}(a) - (m + k + 1)\, q^{m+k}(a)\, p(a) - {}^{2}\Pi_{m+k+1}(a) \quad , \tag{18}$$

where a, $p(a)$, and $q(a)$ have already been defined. The quantity k is the number of check digits required by the Hamming single-error-correcting code. The quantity ${}^{2}\Pi_{n}(a)$ [in analogy to Eq. (12)] is the multiple integral

*The set of possible sequences of n binary digits can be represented by the vertices of a unit cube in a space of n dimensions.[2] The distance between two vertices is defined as the number of binary digits in which the corresponding sequences differ.

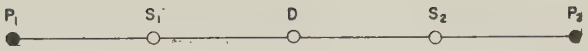


Fig. 1. Configuration of points in message space between two possible transmitted messages P_1 and P_2.

TABLE II

COMPARISON OF HAMMING, WAGNER, AND HAMMING-WAGNER CODES

(a) $\alpha = 1.35$

m	P_U	P_H	P_W	P_{HW}
10	0.093	0.044	0.030	0.037
11	0.111	0.050	0.038	0.043
12	0.110	0.065	0.038	0.057
13	0.128	0.073	0.047	0.064
14	0.146	0.081	0.056	0.071
15	0.165	0.090	0.067	0.079
16	0.183	0.098	0.079	0.087
17	0.202	0.107	0.092	0.096
18	0.220	0.116	0.105	0.104
19	0.239	0.126	0.119	0.113
20	0.257	0.135	0.134	0.122
21	0.275	0.145	0.149	0.132

(b) $\alpha = 1.80$

m	P_U	P_H	P_W	P_{HW}
10	0.0091	0.0016	0.00067	0.00063
11	0.0117	0.0018	0.00095	0.00076
12	0.0109	0.0025	0.00081	0.00107
13	0.0135	0.0029	0.00111	0.00124
14	0.0163	0.0033	0.0015	0.0014
15	0.0193	0.0037	0.0019	0.0016
16	0.0224	0.0042	0.0024	0.0019
17	0.0258	0.0046	0.0030	0.0021
18	0.0292	0.0051	0.0037	0.0024
19	0.0328	0.0057	0.0044	0.0026
20	0.0364	0.0062	0.0052	0.0029
21	0.0401	0.0068	0.0061	0.0032
22	0.0439	0.0074	0.0070	0.0036
23	0.0478	0.0080	0.0080	0.0039
24	0.0518	0.0086	0.0091	0.0043

Values of α are for the Hamming-Wagner code
m = number of message digits

$$ {}^{2}\Pi_n(a) = \frac{n!}{(\sqrt{\pi})^n} \int_o^{\infty} \exp[-(x_n - a)^2]dx_n \int_o^{x_n} \exp[-(x_{n-1} - a)^2]dx_{n-1} \cdots $$

$$ \int_o^{x_4} \exp[-(x_3 - a)^2]\,dx_3 \int_o^{x_3} \exp[-(x_2 + a)^2]dx_2 \int_o^{x_2} \exp[-(x_1 + a)^2]dx_1 \qquad (19) $$

Equation(19) can be reduced to the sum

$$ {}^{2}\Pi_n(a) = \frac{n(n-1)}{2^n} \sum_{i=2}^{n} \binom{n-2}{i-2}(-1)^i\, {}^{2}I_i(a), \quad \text{where} \quad {}^{2}I_i(a) = \frac{2}{\sqrt{\pi}} \int_o^{\infty} [\mathrm{erf}(x-a)]^{i-2} \exp[-(x+a)^2]\,dx \qquad (20) $$

in complete analogy to Eq.(13).

$P_{HW}(1.35)$ and $P_{HW}(1.80)$ are tabulated in TableII for various values of m, together with the corresponding probabilities of error for uncoded, Hamming-coded, and Wagner-coded words. The values of a used in computing P_U, P_H, and P_W are chosen so that all words (message digits plus check digits) have the same duration, as required in a constant-data-rate system. Thus

$$ P_U(a_U) = 1 - q^m(a_U) \quad , \qquad a_U = \sqrt{\frac{m+k+1}{m}}\, a $$

$$ P_H(a_H) = 1 - q^{m+k}(a_H) - (m+k)\, q^{m+k-1}(a_H)\, p(a_H) \quad , \quad a_H = \sqrt{\frac{m+k+1}{m+k}}\, a \qquad (21) $$

$$ P_W(a_W) = 1 - q^{m+1}(a_W) - \Pi_{m+1}(a_W) \quad , \qquad a_W = \sqrt{\frac{m+k+1}{m+1}}\, a $$

TableII shows that for a = 1.35, a very noisy case, the Hamming code becomes better than the Wagner code at m = 21. For a = 1.80, which corresponds to much less noise, the Hamming code surpasses the Wagner code at m = 24. Thus it appears that, starting with some value of m between 25 and 30, the Hamming code is better than the Wagner code anywhere in the significant range of a (neither too little nor too much noise). This happens for the reasons given at the end of Sec. D.

We see from the table that the Hamming-Wagner code is consistently better than the Hamming code; however, the percentage improvement is greater for a = 1.80 than for the noisier case a = 1.35. For a = 1.35, the Hamming-Wagner code is better than the Wagner code for all m > 17; for a = 1.80, the Hamming-Wagner code is better than the Wagner code for all m > 13.* Thus, while the Wagner code is superior to the Hamming code for words of length less than about 20, the Hamming-Wagner code is superior to either of these codes for words of length greater than about 15.** The Hamming-Wagner code works better in low noise than in high noise, because (1) proportionately fewer multiple errors are of order higher than two, and (2) the conditional probability of correcting double errors is higher. Since this conditional probability decreases as m increases, the Hamming-Wagner code gradually becomes less effective, as shown in the next section.

*The Hamming-Wagner code is also better for m = 10 and 11. This anomaly is due to the change in k from 4 to 5 at m = 12.

**The value of m for which one code becomes better than another is somewhat dependent on a. (See Table II.)

G. THE SYLLABIFIED WAGNER CODE

Another multiple-error-correcting code based on the principle of the Wagner code is the syllabified Wagner code, constructed by dividing each word into separately Wagner-coded subwords or syllables. Suppose a word with m message digits is divided into j syllables, each containing $n_i = m_i + 1$ digits, where

$$m = \sum_{i=1}^{j} m_i \quad .$$

Since the probability that a syllable (regarded as a Wagner-coded word) is correct is

$$q^{n_i}(\alpha) + \Pi_{n_i}(\alpha) \quad ,$$

the probability of error for a syllabified-Wagner-coded word is

$$P_{SW}(m_1, m_2, \ldots, m_j) = 1 - \prod_{i=1}^{j} [q^{n_i}(\alpha) + \Pi_{n_i}(\alpha)] \quad , \quad \sum_{i=1}^{j} m_i = m \quad . \qquad (22)$$

It follows at once that for a given number of syllables $P_{SW}(m_1, m_2, \ldots, m_j)$ is smallest when the syllables have equal length (or as nearly equal as possible).

If too few syllables are used, the conditional probability of correction of single errors per syllable is small because the syllables are too long. If too many syllables are used, this conditional probability is small because the large number of check digits leads to a small value of α. (This second effect is partially compensated by increased multiple-error-correction possibilities.) The optimum number of syllables is a compromise between these two effects. This optimum number is not necessarily critical, or for that matter the same for all α. The simple Wagner code (which may be considered a syllabified Wagner code of one syllable) is clearly best for short words. At about m = 14, division into two syllables is better than the simple Wagner code. At m = 30, divisions into three and four syllables are about equally effective, and better than divisions into more or fewer syllables. A syllable length of seven to ten digits seems to be best.

All these points are illustrated in Table III, which compares P_{HW} and P_{SW} for several values of m and $\alpha_{HW} = 1.80$. The table also shows how the syllabified Wagner code finally surpasses the Hamming-Wagner code at about m = 80. As previously mentioned, this is due to the decrease in C_{HW}, the conditional probability that the Hamming-Wagner code corrects double errors, as m increases. This decrease in C_{HW} is also shown in Table II. The formulas used for calculating the P's are the same as those in Eqs.(21) and (22) with the α's related by

$$\alpha_U = \sqrt{\frac{m + k + 1}{m}}\ \alpha_{HW} \quad , \quad \alpha_W = \sqrt{\frac{m + k + 1}{m + 1}}\ \alpha_{HW} \quad , \quad \alpha_{SW} = \sqrt{\frac{m + k + 1}{m + j}}\ \alpha_{HW} \qquad (23)$$

where m is the number of message digits, k the number of check digits, and j the number of syllables.

TABLE III

COMPARISON OF THE HAMMING-WAGNER AND SYLLABIFIED WAGNER CODES

$a_{HW} = 1.80$

m	P_{HW}	C_{HW}	j	P_{SW}
12	0.00107	0.77	1	0.0081
			2	0.0087
14	0.00143	0.75	1	0.00148
			2	0.00144
16	0.00186	0.73	2	0.00228
18	0.00236	0.72	2	0.00322
			3	0.00342
20	0.00292	0.70	2	0.00438
			3	0.00449
22	0.00356	0.68	2	0.00577
			3	0.00570
24	0.00426	0.67	3	0.00700
			4	0.00735
30	0.00730	0.62	3	0.00981
			4	0.00975
			5	0.01006
42	0.0146	0.55	5	0.0200
			6	0.0201
54	0.0244	0.50	6	0.0318
			7	0.0317
72	0.0448	0.43	8	0.0468
90	0.0688	0.38	10	0.0659

j = number of syllables

H. THE REED CODE

We now examine the performance in a constant-data-rate system of the Reed code,[3] an algebraic multiple-error-correcting code. The Reed code is applicable only when the total number of digits in a word is a power of 2. Corresponding to each possible word length, there are only certain possible values of the order to which errors may be corrected. For each of these possible values, the number of message digits is determined. This feature limits the application of the code in communication systems, for the number of message digits in a word (fixed by other considerations) may not correspond to a possible choice in a Reed code. Complete details are given in Reed's paper.[3]

Table IV shows corresponding probabilities of error per word for a Reed three-error-correcting code, the Hamming single-error-correcting code, and no code at all. The formulas for P_U and P_H are the same as those given in Eq. (21); the probability P_R is given by

$$P_R = 1 - \sum_{i=0}^{3} \binom{m+k_R}{i} q^{m+k_R-i}(a_R)\, p^i(a_R) \tag{24}$$

where k_R is the number of Reed check digits, k_H the number of Hamming check digits required for m message digits. The relations required to find the a's used in Table IV are

$$a_U = \sqrt{\frac{m + k_H}{m}}\, a_H$$

$$a_R = \sqrt{\frac{m + k_H}{m + k_R}}\, a_H \tag{25}$$

We see from Table IV that the Reed code outperforms the Hamming code, even for m = 16. Thus the decrease in a produced by the extra check digits of the Reed code is more than compensated by the ability to correct all double and triple errors. The advantage is less marked for smaller a, since in high noise many more errors of order greater than three are introduced by the shortening of the digit length.

In Table V, the Reed code is compared for three of its allowed values of m with the best of the Wagner codes. The probability of error for uncoded words is given for reference.

We see that the Wagner-type codes can compete with the Reed code in high noise. As the noise decreases or m increases, the Reed code increases its advantage. It is clear that for ordinary communication purposes, the Reed code would be better for long words than any of the previously considered codes if the restriction on the allowed values of m could be removed.

TABLE IV

PROBABILITIES OF ERROR FOR UNCODED, HAMMING-CODED, AND REED-CODED WORDS

m	a_H	P_U	P_H	P_R
16	1.5	0.114	0.049	0.047
16	2.0	0.00953	0.00112	0.00042
42	1.5	0.389	0.195	0.163
42	2.0	0.0512	0.0057	0.0011
99	1.5	0.754	0.538	0.449
99	2.0	0.1558	0.0259	0.0041
219	1.5	0.967	0.899	0.839
219	2.0	0.354	0.100	0.018

TABLE V

COMPARISON OF REED CODE WITH HAMMING-WAGNER AND SYLLABIFIED WAGNER CODES					
m	a_{HW}		P_U	P_{HW}	P_R
16	1.80		0.0224	0.0019	0.0022
42	1.80		0.1181	0.0146	0.0097
m	a_{SW}	j	P_U	P_{SW}	P_R
99	1.50	10	0.726	0.359	0.403
99	2.00	10	0.1379	0.0151	0.0029

I. SUMMARY AND CONCLUSIONS

We have considered the use of several types of binary codes in communication systems, making the following assumptions:–

(1) The system transmits sequences of binary digits known as words. If any digit is altered, the information carried by a word is lost. Thus, by definition, sequences obtained by combining words are not themselves words.

(2) The transmitted digits are one of two electrical signals of bandwidth W and duration T. They have equal energies and equal a priori probabilities.

(3) The entire coded word must be transmitted in a given time, regardless of the number of code digits required to check the message digits. (Assumption of constant data-rate.)

(4) The transmitted digits are corrupted by the addition of white Gaussian noise. They are determined by choosing the larger of two independent and normally distributed correlator outputs. The time-bandwidth product, TW, of the transmitted signals is $>> 1$, so that when the signal length is changed to accommodate different numbers of check digits, the signal-to-noise ratio of the correlator difference voltage is proportional to the square root of the signal length.[5] (Actually, the signal-to-noise ratio is proportional to $\sqrt{TW}$, but we assume that W is not changed, an assumption that requires $TW >> 1$.)

By the best code (of those we consider) for a given word length and channel noise, we mean that for which the probability of error per word is smallest (under the assumption of constant data-rate). We have considered the following systematic codes: – (1) the Hamming single-error-correcting code, (2) the Wagner code, (3) the Hamming-Wagner code, (4) the syllabified Wagner code, and (5) the Reed multiple-error-correcting codes. (The Wagner, Hamming-Wagner, and syllabified Wagner codes are introduced in this paper.) For short words (m < about 15) we find that the Wagner code is best in the range of interest (neither too little nor too much noise). As m increases, the Wagner code is surpassed by both the Hamming-Wagner code and a syllabified Wagner code of two syllables.* For values of m < about 80, all syllabified Wagner codes are

*The Hamming code surpasses the Wagner code for m about 20, but is always inferior to the Hamming-Wagner code.

inferior to the Hamming-Wagner code. For larger m, the conditional probability that double errors are corrected by the Hamming-Wagner code has fallen sufficiently so that a syllabified Wagner code is better. Thus, were it not for the Reed code (which is only applicable for a few word lengths), we could say that the Wagner code is best for short words, the Hamming-Wagner code for medium length and long words, and the syllabified Wagner code for very long words. However, the Reed code outperforms the Hamming-Wagner code at m = 42 and the syllabified Wagner code at m = 99 (except in excessively high noise), showing that for large m there is no substitute for an algebraic multiple-error-correcting code. We can safely say that if the Reed code can be generalized to apply to any number of message digits, it will be the best code except for short words. This assumes that the proportion of check to message digits turns out to be comparable to that of the present Reed code.

The numerical work reported here was done by Mrs. Elizabeth Munro.

REFERENCES

1. C.E. Shannon, Bell System Tech. Jour. 30, 50-64 (January 1951).
2. R.W. Hamming, Bell System Tech. Jour. 29, 147-160 (April 1950).
3. I.S. Reed, "A Class of Multiple-Error-Correcting Codes and the Decoding Scheme," Technical Report No. 44, Lincoln Laboratory, M.I.T. (9 October 1953).
4. P.M. Woodward and I.L. Davies, Proc. I.E.E. 99, 111, 37 (March 1952).
5. W.B. Davenport, Jr., R.A. Johnson and D. Middleton, Jour. Appl. Phys. 23, 4, 377-388 (April 1952).

INFORMATION, ORGANIZATION AND SYSTEMS

Jerome Rothstein
Columbia University
New York, N.Y.

The object of this paper is to develop and apply a mathematical concept of organization and of systems. It is very closely related to the information concept and provides the link whereby the theorems of communication theory become generalized and applicable to systems in general. Brief applications are given to system reliability, the significance of organization theory for circuit design, and production and quality control for a systems viewpoint.

The Nature of Organization

Intuitively, one equates disorganization to chaos. This suggests the possibility of measuring quantity of organization by the amount of information required to specify an organization in terms of its unorganized components, or by the entropy increase occurring when the organization is dissolved. The two approaches coincide, and organization can be regarded as a generalization of the entropy concept, just as information is.

Consider a set of elements, each associated with a set of alternatives. It is unnecessary to specify the nature of elements or associated alternatives for the mathematical theory, but it may be helpful to keep concrete examples in mind, e.g., a message source or a physical system as the element, and the ensemble of possible messages or the set of operationally distinguishable states of the system respectively as the associated sets of alternatives. The elements need not be of similar natures, nor need the various alternatives associated with a given element have more in common than that association.

Call each set of alternatives associated with an element a space. Restrict these spaces to the class of measure spaces on which a probability measure and thus entropies can be defined for subsets of each one. Consider the measure space formed by taking the Cartesian or direct product of all these spaces. A "point" in the product space can be looked at as a "vector," each component being a point in the space of some element. Probabilities and entropies can be defined for subsets of the product space. The set of elements will be said to be unorganized or have zero organization (synonyms: mutually independent, statistically independent, uncoupled, not in communication with each other, unconnected, etc.) if the entropy of a set of points in the product space is the sum of the entropies of the corresponding sets of points in each of the spaces associated with an element. This is clearly the case of maximum set entropy in product space for specified sets in each element space.

The essence of organizing a set of elements resides precisely in the fact that elements do influence each other. Synonyms for organized are: coupled, linked, correlated, connected, in communication with each other, coordinated, interacting, etc. It is in these couplings, correlations, or constraints that the organization consists. A measure of amount of organization entailed by the couplings is the concomitant reduction of entropy in product space compared to that calculated for the unorganized state of the same elements. Organization, like information, is thus a negative entropy. It is maximal for perfect correlation between elements, i.e., functional rather than stochastic relations between them. Consider the case of two elements (generalization to a finite number is simple), each of whose spaces can be represented by a line segment over which a probability distribution is given. The product space is then a rectangle with the segments as sides. For zero organization, assuming probability densities are defined on the segments, a probability density is defined over the rectangle and is simply the product of the segmental densities. For maximal organization, the domain of non-vanishing values of the distribution in product space shrinks to a set of zero planar measure, e.g., a curve. The segment distributions and this one do not differ essentially, as each is a one to one map of the others. In general, the segment distributions are marginal distributions corresponding to integrating out one of the two variates. For organization between zero and maximal, the probability density in the rectangle is peaked in some regions and lowered in others compared to the product of the segmental densities, with maximal organization a kind of delta function limit of this. More explicitly, if the probability density in the rectangle is $p(x,y)$, the marginal distributions are

$$p(x) = \int p(x,y)dy \tag{1}$$

$$p(y) = \int p(x,y)dx \tag{2}$$

with the corresponding entropies (using Shannon's notation) given by

$$H(x,y) = -\int p(x,y) \log p(x,y)\, dx\, dy \tag{3}$$

$$H(x) = -\int p(x) \log p(x)\, dx \tag{4}$$

$$H(y) = -\int p(y) \log p(y)\, dy \tag{5}$$

the amount of organization, ΔH, is

$$\Delta H = H(x) + H(y) - H(x,y). \tag{6}$$

In terms of conditional entropies, $H_x(y)$, $H_y(x)$, and the known relations

$$\begin{aligned} H(x,y) &= H(x) + H_x(y) \\ &= H(y) + H_y(x), \end{aligned} \tag{7}$$

this can be rewritten in the forms

$$\begin{aligned} \Delta H &= H(x,y) - H_x(y) - H_y(x) \\ &= H(y) - H_x(y) \\ &= H(x) - H_y(x) \end{aligned} \tag{8}$$

From any of these expressions for ΔH, it follows immediately that ΔH vanishes for statistical independence of the two variates, and is a maximum equal to either univariate entropy, for functional connection between them.

Systems

Intimately connected with the concept of organization is that of a system, defined as an organization with a function, i.e., it couples two (or more) ensembles of interest. Synonyms for function are: task, program, behavior pattern, stimulus-response, input-output, etc. System plus function constitute an organization as defined earlier. Reasons for introducing a separate definition are practical, e.g., (a) the function is often specified in advance and not under the control of the system designer, whereas (b) design of a system to realize the given function is the center of interest and under the control of the system designer, (c) simplifications often result if part of a complex organization whose internal interactions are much stronger than its interactions with other parts is treated as an entity in itself, (d) dealing with system and function often requires entirely different techniques and one can often be treated almost independently of the others, (e) except for requirements that the system be optimum in some sense or senses, e.g., economical, dependable, etc., the particular elements and associated ensembles of alternatives are not of interest in themselves but only as means to an end, viz. performance of the desired function. The function of the system can be simply described as organizing a product space, e.g., the product of input and output spaces. A communication system is a system in this sense, its function to couple an ensemble of messages at a source to an ensemble of messages at a destination. Similarly, a physical theory is a system for predicting - it couples states at a future time (output) say, with initial states (input). Alternatively expressed, theory organizes observation. Other examples from many are strategies, which organize one's play (cut down the ensemble of possible moves in a given position; input is the position before moving, output is the position after moving), manufacturing systems (input of raw materials, output of finished items, manufacturing tolerances correspond to permissable noise level, etc.), transportation systems, automatic control systems, etc.

Consider a system with input space (x), output space (y), and marginal entropies H(x), H(y), H(x,y). The amount of organization introduced by the system is clearly given by (6) or (8). The same equations hold if entropy is replaced by entropy rate. For the special case of (x), the message space at a source, and (y), the message space at a destination, these are familiar equations for rate of transmission of information. In general, they give the rate (usually denoted by R) at which the system organizes the product space. The maximum rate for all input ensembles is called the system capacity and can be expressed as

$$C = \lim_{T\to\infty} \operatorname{Max}_{p(x)} \frac{1}{T} \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \, dx \, dy \tag{9}$$

C is the channel capacity for a communication system.

It should be noted that time rates are not the only ones of interest, though they are probably most important. Storage capacity of a given medium for information is another. For a magnetic tape, for example, T would represent length of tape, and C could be measured in bits/cm. Similarly, T would be area for information stored on photographic film, etc. Different capacities are often related when systems are formed, as in the case of transmission of information stored on a magnetic tape. If the information is read off the tape moving at speed v, and it has storage capacity C (per unit length), then the minimum channel capacity required to avoid information loss is vC. Factors analogous to v come up (derivatives, Jacobians) when different media are coupled.

Before returning to the general theory, it seems advisable to examine briefly a system other than a communication system, e.g., a production unit of a manufacturing enterprise. The entire enterprise is an organization of which the production unit is an element. In its interaction with the rest of the organization, it has input and output of material and information. It is thus itself an organization with a function, or a system. Its input consists of the raw or partly processed material on which it operates and the input and output specifications. The output consists of the partly finished item to be passed on further in the organization for additional treatment or the end item itself. Its capacity (here, productive capacity) is the maximum rate at which it can produce acceptable output items. The input and output ensembles are sets of mathematical entities, e.g., numbers, serving to describe the incoming and outgoing objects with desired exactness. They often consist of the results of measurements made on the object, go - no go indications, and the like. Each object, incoming and outgoing, corresponds to a "point" in a space (usually discrete). The statistical population consisting of the

totality of these objects defines input and output distributions. The specifications require that the object points fall in certain regions. If the entire output falls within the output specification, the production unit is reliable, the "message" (i.e., specification) has been transmitted to the "destination" (output object) through the "channel" (production unit). When the "message" is garbled by "noise," i.e., the specification limits are exceeded, the object is rejected on inspection. Just as in a communication system where one strives not to make sure that all messages are received correctly but rather to keep errors below some acceptable level, so in manufacturing processes, one aims not at making all items identical but rather to keep them within acceptable tolerance limits. It can be seen that statistical quality control in production and statistical communication theory are indeed closely related. Furthermore, the general theorems of communication theory apply to all systems; little difficulty arises in general in specifying input and output ensembles.

Returning now to the general theory, the rate of organizing the input-output ensemble is clearly decreased by any weakening of the coupling between them. Errors, unreliability, noise, breaking of a physical connection included in the coupling means, etc., tend to increase Hy(x) in the expression for rate

$$R = H(x) - Hy(x). \tag{10}$$

If the error statistics are independent of those of the input and output ensembles (corresponding in communication to having noise statistics independent of what messages are chosen), an error entropy, analogous to noise entropy in communication theory, can be defined. In many important situations, this entropy, denoted by H(N), includes all of the deviation from maximum organization, in which case, one can write

$$R = H(x) - H(N). \tag{11}$$

This can be extended to yield a generalization of Shannon's theorems on the maximum rate of transmission of information in terms of available signal and noise powers, if a suitable superposition principle for signal and noise is satisfied. For the magnetic tape earlier discussed, this is the case with noise power replaced by minimum significant variation in intensity of magnetization and maximum signal power by saturation tape magnetization. The techniques of wave-form analysis also apply, with time domain replaced by space domain, frequency by wave-number (reciprocal of wave-length). All the theorems carry over and in particular, with suitable coding, a storage capacity arbitrarily close to the theoretical maximum can be obtained.

Concluding Remarks

The usefulness of an analogy resides in making an unfamiliar field come under previously known concepts, in making previous knowledge applicable to new situations. Information theory, or cybernetics, has been doing precisely this, and the present paper attempts to carry the process further. The organization concept appears to be a generalization of the information concept in that it (a) in no way depends on the asymmetry between transmitter and receiver implicitly assumed in communication theory; (b) handles any number of interacting ensembles, not just the conventional pair of communication theory; (c) includes information as a special case. Examples based on circuit considerations where the new concepts are applicable will now be given.

In the field of circuit design, one might expect information theory to have considerable impact, but this has not been the case. The reasons appear to be (a) that the components of circuit theory are either taken to be noiseless, so that each function is merely a one-to-one mapping, trivial from the information-theoretic viewpoint; (b) many special methods of reducing noise (shock mounts, shielding, ruggedized tubes, etc.) are effective without the use of complicated theory; (c) statistical filtering and the like is justified in relatively few cases.

One would expect information and organization theory to be of real importance in cases where the component is truly an element of an organization, with output ensemble specified only statistically by its input. The two chief cases of this appear to be (a) when the individual circuit includes elements whose behavior is specified statistically, and (b) when the individual circuit is determinate but one of a statistical population of circuits made to the same specifications, whereby the input to a given element reflects the statistics of elements feeding it, and its output reflects the additional effects of its own probability distribution over a set of performance-determining parameters. It is clear that the two cases are closely related mathematically, but they may be appropriate to entirely different physical situations.

Case (a) would include, for example, complex electronic systems like computers, where reliability (i.e., elimination of "generalized" noise) becomes important even for individual elements. The controlled use of "redundancy" (i.e., channel capacity in excess of that required to transmit information at a specified rate in the absence of noise) to reduce the effects of noise is clearly applicable to increasing the reliability of the computer. Self-checking codes are precisely this. In general, a similar employment of additional system capacity can increase system reliability, another example being a decrease in reject rate sometimes achieved by establishing inspection points at intermediate stages of a production operation. Case (b) comes up when one is faced with the problem of designing mass-produced equipment to some performance specification with preassigned tolerances specified for the components. When the number of components is large with ordinary design methods, even moderate performance specifications can make unreasonable demands on components. Here information or organization theory may well be crucial in design, for the components are "noise" sources which in their totality blot out the "message" (falling within performance limits). "Redundant" design appears mandatory, and can be viewed as a sophisticated version of the familiar "factor of safety."

AN INFORMATION-THEORETICAL MODEL OF ORGANIZATIONS

Manfred Kochen
Institute for Advanced Study and Paul Rosenberg Associates
Princeton, N. J. Mt. Vernon, N. Y.

Introduction

An organization can be treated as a set of component members which, as a group, is capable of performing functions which the individuals are not. This property is due to the ability of each member to influence, and be influenced by every other one in the system.

It is meaningful to consider only those organizations which can be defined operationally, relative to a given observer. The observer obtains information about the system he is describing by interacting with the system much in the same manner in which the components of the system interact with one another. Just as the above-mentioned observer-system pair may be viewed as an organized system in itself within a larger environment, each member of the originial system, and any collection of these, can be viewed as subsystems also. The observer describes the system upon which attention is focussed according to a task or function, and the efficiency with which the system performs it. The function of the entire group as well as that of each member is assumed to be physically observable and measurable by the observer, although not necessarily by the members.

The basic notion which is exploited here is that every potential member of an organization acts so as to maximize the value resulting from his action to him. Organizations are assumed to exist, because the value of belonging to the group is greater to each member than of not belonging. This holds clearly also in cases where members are constrained or coerced to be in the group, in which case the value to these members of not belonging is more negative than that of belonging. The potentially most highly organized systems are those in which there exist states of the system for which the values to every member are most nearly all maximized.

Although a system may be highly organized potentially, the potential may be far from realized, in which case the efficiency is said to be low. It will be seen that this efficiency may be measured in terms of the uncertainty of each member about the actions of the others; that this uncertainty is best removed by efficient coding procedures, which admit of the maximum of intercommunication within the group, considering limited storage capacities.

It is the purpose of this paper to make the above concepts more precise; to establish a formal mathematical model based upon five axioms which define the members, and characterize their permitted behaviour; to apply the well known results of information theory in order to determine which modes of communication and action should be, can be, and are most likely to be evolved for each member to maximize his value function; these being assumed as known to the observer to whom these questions are directed.

The Formal Model

Axiom 1:

A set S_i is assigned to each member P_i of the group $P = (P_1, P_2, \ldots, P_N)$. In the duration of time Δt, P_i may choose one of the elements of S_i, denoted as s_i by the observer who is outside of the system.

The specification of the set S_i is the task of this observer also, and the set may be non-denumerable, as, say the set of all real numbers in the closed interval $[0,1]$, countable, such as the set of all integers, or finite, depending upon the observer. In most of the discussion to follow, all the S_i will be assumed to consist of m elements, although not necessarily the same ones for each S_i.

The notion of choice is left undefined, but it is assumed that the result of each choice has an effect on P_i, $i = 1, \ldots, N$ and on the observer, which is capable of being recorded in coded form:

Axiom 2:

A set of "code symbols" composed of binary elements (zeroes and ones), S_{ji}, is assigned to each P_i with reference to P_j. The effect of a choice by P_j is the selection of some element from S_{ji}, denoted as s_{ji} to P_i, the selection being made by P_i. That is, P_i encodes as s_{ji} what the outside observer calls s_j.

Crudely speaking, when P_j chooses and performs action s_j, P_i "interprets" this action as s_{ji}. The axiom essentially states that each member is able to receive the effects of actions made by other members, and make statements about these actions. The actions may be expressed statements themselves.

It is not necessary that S_{ji} and S_j be in one-to-one correspondence, and it is, in fact, this, which contributes to P_i's uncertainty about P_j.

Axiom 3:

A set function, v_i , which maps each state of the system as encoded by P_i, $\mathcal{G}_i = (s_{1i},\ldots,s_{Ni})$ into an element of an image set, T_i, is assigned to each P_i. The set T_i is assumed to be partially ordered.

The function v_i is to be interpreted as defining the value of some states of the system to P_i relative to some or all other states. If the ordering is total, then for any two states, one is either preferred, considered indifferent to, or inferior to the other. In some cases, the assumptions on T_i may be strengthened; open sets and neighborhoods may be defined in it, so that it is possible to say that one state of the system has a value to P_i which is "close" to that of another state. Further, T_i may be considered a metric space, with some of the properties of number sets. The choice of the specifications of the T_i as well as the v_i depends upon the observer's knowledge of the system, and is known, a priori, to him only.

A state $\mathcal{G}_i$ is said to be a maximal state for P_i if it is preferred to every other state known to P_i. There need not be a unique maximal state for every v_i. In case S_i i=1,...,N and T_i are topological spaces, v_i is defined to be continuous; then, even though a maximal state may not always exist, a state for which v_i takes on its least upper bound (supremum) always exists.

If it is supposed that S_i and S_{ji} are in 1-1 correspondence for all i, j, and there exists a state which is a maximal state for all the P_i, the organization is potentially perfect.

Axiom 4:

P_i chooses that element from S_i, which, according to his data about the ordering on T_i, maximizes his expected value.

If it is temporarily assumed that v_i and the frequencies of the various possible choices and states of the system are known to P_i, Axiom 4 is to be interpreted in the sense of game theory.[1,2] There are, of course, several other ways of formulating axiom 4, such as the minimax principle used in game theory, the Bayes criterion, Hurwitz criteria, etc. Here the observer again specifies the criterion according to which the members make their choices.

Let the T_i be number sets. The quantity

$$\xi = \operatorname*{Max}_{\mathcal{G}} \frac{1}{N} \sum_{i=1}^{N} \xi_i(\mathcal{G}) \qquad \text{where} \qquad \xi_i(\mathcal{G}) = \frac{V_i(\mathcal{G})}{\operatorname{Max} V_i}$$

may be taken as a measure of the order of the organization. It is easily seen that if the system is potentially perfect, all the ξ_i and hence $\xi(\bar{\mathcal{G}})$ is 1, where $\bar{\mathcal{G}}$ is the maximal state. A value of 0 for ξ is interpreted to mean that the system is completely disorganized. It is quite possible that for certain kinds of organizations, a lower limit on ξ can be determined, which is to serve as a disintegration threshold.

That there exist value functions in systems for which a state satisfying all the members does not exist can be demonstrated by the voting paradox. Let N=3 and consider 3 distinct states of the system, $\mathcal{G}_1\,\mathcal{G}_2\,\mathcal{G}_3$. In the following table, + denotes a high value, and - a low one. If a group value function is defined as + when two or three of the three members assign the value + to a preference, it is easily seen that the group value is not consistent in the sense of transitivity. > denotes "is preferred to".

	$\mathcal{G}_1 > \mathcal{G}_2$	$\mathcal{G}_2 > \mathcal{G}_3$	$\mathcal{G}_1 > \mathcal{G}_3$
P_1	+	+	+
P_2	-	+	-
P_3	+	-	-
P	+	-	-

Axiom 5:

P_i may store up to c_i binary digits, and may change at most r_i bits per time unit Δt . These quantities are important in determining the amount of information P_i may store if he encodes it properly, and the number of other members with which he may be in contact per time interval.

It is not specified what P_i may store, but under certain conditions there is always an optimal set of quantities and an encoding procedure, in the sense that P_i's value will be as large as possible. There are also optimal communication networks which will lead to the realization of the ξ of the group.

Data

The time interval Δt, which has been referred to above, and which was defined as the duration in which only one choice from S_i may be made, shall henceforth serve as a time unit. Let t be an integer which denotes the number of time units which have elapsed since the observer began to observe the system. It is now assumed that the following data is available to P_i, i=1,...,N, at time t:

1. The t states through which the system, as encoded and stored by P_i, has passed.
2. The t relative values associated with each state.

Clearly, only so much of this information is actually available to P_i as can be suitably stored by him. The data, as described above, can be summarized in the form of a rectangular matrix, at most Nxt. $D_{it} = (d_{i1},d_{i2},\ldots,d_{it})$, where d_{it} is the transpose of the row vector $(s_{1i,t},\ldots,s_{Ni,t}, v_{it})$. Because of P_i's limited storage capacity, all the entries in this matrix need not be filled. All the actions and decisions P_i is able to perform must be based on this data only.

All the N data matrices, the knowledge of the v_i, etc. are, of course, available to the observer who describes the system also.

Variation with time

At t=0, which might be called "a priori", P_i has no basis whatever for any action; nor does any other member. During the time interval $(0, \Delta t)$ P_i may receive and store the choices of as many as 2^{c_i} members, including himself, such that no two of these correspond to the same sequence of bits. During the next time interval, P_i receives a message about a choice from, say P_j. P_i first compares the message with the one he has received from P_j (if any) during the previous time interval. If the two messages are the same, it is unnecessary to waste storage by recording the second message; it is sufficient to indicate the fact that there are now two members in the category which was established during the first period. If the second message differs from the first, a new category in the set S_{ji} is established. In this way, the sets S_{ji} for all j and i are built up in time; they may even be called "ensembles", because frequencies are associated with the elements.

Consider the relationship of P_i influencing P_j. This may be defined operationally relative to P_k as the condition in which P_i's choice is followed by a certain choice of P_j with a frequency greater than could be attributed to chance alone. It would be reasonable to require that if P_i influences P_j strongly relative to P_k, P_i should also influence P_j strongly relative to any other $P_{k'}$, who has information about the pair. The consequences of this requirement have yet to be explored.

Examples

In the following examples, N=2. The extensions to N> 2 are not difficult to imagine.

Example 1. Defense team.

The purpose of the pair, as determined by the observer, is to destroy an opponent. Let S_1 consist of the following two elements:

$s_1^{(1)}$: to act as decoy; to sacrifice, drawing the opponent's entire attention.

$s_1^{(2)}$: to strike at opponent, with assurance of destroying him.

S_2 consists of the analogous two alternatives, plus a third:

$s_2^{(3)}$: to feign or bluff, neither striking nor sacrificing.

Define the codes as follows:

$$s_{11}^{(1)} = d \qquad s_{11}^{(2)} = k \qquad\qquad s_{22}^{(1)} = d' \qquad s_{22}^{(2)} = k' \qquad s_{22}^{(3)} = f'$$

$$s_{21}^{(1)} = d \qquad s_{21}^{(2)} = k \qquad s_{21}^{(3)} = d \qquad\qquad s_{12}^{(1)} = d' \qquad s_{12}^{(2)} = k'$$

$$v_1(d,d) = t \qquad v_1(k,d) = g \qquad\qquad v_2(d',d') = t' \qquad v_2(k',d') = g' \qquad v_2(f',d') = b'$$

$$v_1(d,k) = b \qquad v_1(k,k) = g \qquad\qquad v_2(d',k') = b' \qquad v_2(k',k') = g' \qquad v_2(f',k') = e'$$

In this example, the letters t,b,g,e may be thought of as representing the values: terrible, bad, good, excellent, as interpreted by P_1; these letters with primes denote the analogous values to P_2. For instance, $v_2(d',d') = t'$ states that P_2 associates the value "terrible" to what he interprets as a simultaneous choice by himself and P_1 to sacrifice. This example illustrates the importance of communication to organizational efficiency; also one of the ways in which communication may fail, since P_1 confuses the choices of f' and d' by P_2 in this example.

Example 2. Duet of singers.

Suppose that the objective of this duet is a harmonious rendition of a musical composition for two voices. $S_1(t)$ is the collection of all the possible sounds which P_1 could choose to make during the period $(t, t+\Delta t)$, and $S_2(t)$ is similarly defined for P_2. S_{21} is the set of all the possible sounds which P_1 has heard P_2 make, so that any sound which P_2 makes will be classified as some element of S_{21}. S_{12} is the analogous set for P_1. The value functions, $v_1(s_{11}, s_{21})$ and $v_2(s_{22}, s_{12})$ depend upon the critical discrimination and musical abilities of P_1 and P_2 respectively. An experiment can easily be designed in which P_1 and P_2 can rate the value of any particular tone combination (state of the system). It is intuitively clear in this example how each singer maximizes his value function, and what factors determine the extent to which the pair can achieve the objective.

Example 3. Production team.

Let P_1 be a worker, and P_2 his boss. It is quite unnecessary to provide an objective for the pair other than that each tries to maximize his value function (the value to himself). This was the case in

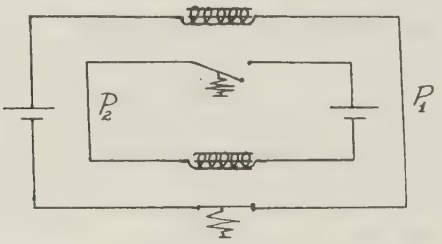
the previous two examples also, the objectives having been stated for clarity. Let S_1 be the set of all the possible levels of intensity at which P_1 is capable of working. S_{11} is a scale by which P_1 measures or "judges" this quantity. S_{21} is a scale, subjective for P_2, by which P_2 can measure the same quantity. That is, an element of S_{21} represents the boss's interpretation of how hard the worker works. Let S_2 be the set of all possible rewards to the worker of which the boss is theoretically capable. S_{22} is the scale by which P_2 measures this, and S_{12} is the scale by which P_1 measures it. $v_1(s_{11}, s_{21})$ is the value to P_1, in P_1's subjective value scale, if P_1 chooses to work with intensity s_{11}, and if P_2 chooses to reward P_1 to extent s_{21}, both measured according to P_1's scale for these. v_2 is similarly defined.

This example illustrates how the selection of an alternative may differ from a statement or an interpretation about this selection. Communication proceeds essentially by making statements about statements about statements etc.; more precisely, this amounts to a time-varying coding procedure. The significance of the order in which the members make their choices depends upon the extent to which communication throughout the whole group prevails. The effective duration and speed of the evolution of organizations will also depend in part upon the extent of communication, which is discussed in the next section.

Example 4. A Learning Experiment.

In the Bush-Mosteller[3] model of learning, applied to a rat in a T-maze, P_1 is the experimenter, P_2 the rat. S_1 consists of the two alternatives: reward, non-reward; these are encoded by P_1 as, say E_1 and E_2, forming the set S_{11}. S_2 is the pair of the rat's alternatives: to turn right, or to turn left; P_1 encodes these as A_1 and A_2, forming the set S_{21}. It is not known how the rat encoded his own choices or those of P_1, but if the outside observer understood the rat sufficiently, it is assumed that the code could be determined. The value function for the rat may be surmised quite easily; for the experimenter, the value function depends upon his expectations and his prior knowledge of the behaviour of rats in such situations. It is noteworthy that in such an experiment as this one, both the rat and the experimenter "learn": acquire information by removing uncertainty about the other's behaviour.

Example 5. A Mechanistic Physical System.



The two magnetically coupled circuits shown in the figure are offered as a final example of an extreme kind of organization. P_1 is the outer circuit, P_2 the inner. S_1 is the pair of possible positions of the switch in P_1, and S_2 the corresponding pair for P_2. S_{11} is the pair of physical states which accompany each position of the switch of P_1; these two states of P_1 might be: the current in P_1 being below a fixed level, and above. S_{21} is a pair of physical states of P_1 which accompany each position of the switch in P_2; these might be the force acting on the spring holding P_1's switch in place with a value greater and less than a fixed number.

In the figure, the springs exert forces on the switches so as to keep them in the positions shown when no current flows in either circuit. When the lower switch is closed, there is current in P_1 and a field is created by the upper electromagnet, closing the upper switch. This, in turn, activates the lower electromagnet, opening the lower switch, and causing the upper switch to open again after a slight delay. It is essentially a double feedback system, and will oscillate.

The value functions are identical for P_1 and P_2, and may be interpreted as follows: there are only four possible states for P_1 and P_2; (a state for the entire system in this case means a pair of states, one from each member) the value to P_1 or P_2 of a state which is consistent with Maxwell's equations and Hooke's law (if it applies) is "acceptable", i.e. high; the value of a state which is inconsistent with these physical laws is low, "unacceptable".

Communications Within the System

Introduction of Information Theory.

The relationship between any two members of the organizations as thus far discussed may be regarded as that of receiver to source in a communications system. Clearly, the roles of receiver and source may be interchanged, and a different channel is, in general, obtained. The channel capacity is given by axiom 5. If the value functions are specified by the observer, the chief and fundamental problem in the description of organizations which remains is to describe the extent of communication within the group. In order to apply Shannon's fundamental results on information and coding[4], it is necessary to define a probability distribution on the set of alternatives.

The measure theoretic approach, upon which Shannon's definition of uncertainty or entropy is based, is very useful in establishing limit theorems, but the Borel fields and the measures defined on these must be known a priori or must be determined by physical methods before the results can be applied to the systems treated here. In cases where either no assumptions at all are warranted, or where those assumptions that may be made about the a priori measures and sample spaces are of such a nature as not to be amenable to analytic treatment, an approach along the lines of distribution-free methods seems relevant. That is, the uncertainty which each member experiences about the possible moves another member might make, must be defined entirely in terms of the data which is stored by that individual.

Consider any two members of P, P_i and P_j. Let $s_{ji,t}$ denote that element from the set S_{ji} which was recorded by P_i during the t th time unit, or observation period. Since P_j is assumed able to make at most one choice during this period, $s_{ji,t}$ is a particular value of the variable s_{ji}. It has already been mentioned that P_i is assumed to be able to decide whether $s_{ji,t}$ and $s_{ji,t'}$ are equal or not. Equality, which in this case connotes the ability of P_i to recognize the recurrence of an s_{ji}, can be taken as an undefined term, as part of axiom 2. It is further presupposed that each P_i is able to count the number of such recurrences and to operate with these numbers according to Peano's axioms.[5] It is only in this manner that frequencies can be meaningfully defined for P_i, and used to utilize his storage capacity and transmission rate best. It seems as though most of the essential features of computers had to be postulated in defining each P_i.

To simulate Shannon's definition of unconditional entropy, it is expedient to define the frequency with which s_{ji} has occurred during the first t trials, denoted by $f_{ji,t}(s_{ji})$, as the ratio of the number of times that s_{ji} has recurred up to time t, to t. It is noteworthy that this is the first point in this model where numbers must be used, the f's being defined on the field of rationals, the properties of which may be used in analysis. The formula

$$U_{ji,t} = -\sum_{s_{ji} \in S_{ji}} f_{ji,t}(s_{ji}) \log_2 f_{ji,t}(s_{ji}) \tag{1}$$

expresses the uncertainty experienced by P_i, on the basis of information in the data matrix D_{it}, about which element of set S_{ji} P_j will choose. Any change in the frequency distribution towards greater concentration decreases $U_{ji,t}$, and the difference in the uncertainties may be taken as a measure of the amount of information gained. P_i may control the change in these frequencies by repartitioning the set S_{ji}.

An expression for the conditional uncertainty experienced by P_i, on the basis of information in D_{it}, about which element of S_{ji} P_j will choose in response to the succession of states which occurred at times t-1, t-2,..., t-T, is given by

$$U_{ji,t}(\mathcal{G}_{t-1},\mathcal{G}_{t-2},,,,\mathcal{G}_{t-T}) = -\sum_{s_{ji} \in S_{ji}} f_{ji,t}(s_{ji}/\mathcal{G}_{t-1},\mathcal{G}_{t-2},..,\mathcal{G}_{t-T}) \log_2 f_{ji,t}(s_{ji}/\mathcal{G}_{t-1},...,\mathcal{G}_{t-T}) \tag{2}$$

where $f_{ji,t}(s_{ji}/\mathcal{G}_{t-1},...,\mathcal{G}_{t-T})$ is the conditional frequency of s_{ji} given $\mathcal{G}_{t-1},\mathcal{G}_{t-2},...,\mathcal{G}_{t-T}$. If the set S_{ji} remains unchanged in time, the frequencies and the uncertainties may stabilize. Before stabilization has taken place, however, $U_{ji,t}$ behaves like a random variable, which is also rational-valued. If values of s_{ji} occur which have not occurred previously, S_{ji} is altered in that the number of elements or categories in it is increased by one. Thus, although no uncertainty is removed in such a case, it is reasonable to assign to P_i a gain of information of the amount of the number of bits required to store the new element.

Storage and Channel Capacities.

Suppose that each of the sets S_{ji} consists of m elements at time t. If only elements of these sets which have previously occurred appear, the system may be found in any one of m^N states. Hence there are m^{TN} ways in which the condition appearing in equation (2) can be realized, and accordingly many items must be stored. The case of unconditional entropy, in which no past state of the entire system is needed, can be considered as the special case T = 0, by definition. By time t, assumed much greater than T, P_i need have storage of at most I bits, being required to store at least the following quantities:

N log m bits which P_i may receive and classify in $S_{1i},...,S_{Ni}$ during period t.

N log m bits with which the above must be compared in order to classify. These are permanently stored.

2NT log m bits as standards for comparison and for receiving the particular states $\mathcal{G}_{t-1},\mathcal{G}_{t-2},...,\mathcal{G}_{t-T}$ which occurred for the last T observation periods. Half of these bits are stored permanently.

N(T+1) log m bits to store the value to P_i from the present and the last T states. If it is assumed that the same value corresponds to two recurring combination of states, this number is multiplied by the number of such combinations which have occurred up to t. At most t-T such combinations may have occurred, observation having started at t > T.

Nm log (t - T) bits to store the frequencies which appear in formula (2).

$$I = N(T+1)(2+t-T) \log m + Nm \log (t-T) \text{ bits} \tag{3}$$

The logarithms in formula (3) and other equations where not explicitly stated shall be understood to be to the base 2. In equation (3) the quantities m and T may both be functions of the time t and the individual whose storage capacity is defined, i. N, the number of members, might be regarded as a function of time, t, in general, since some members may drop out of the organization, and some could conceivably enter also.

It will generally be a rare occasion which requires as many as I bits of storage at time t. Shannon's fundamental theorem may be applied as follows: Regard the entire system, P, acting for T time units as the source, the output of which are the possible states at t which follow the sequence of fixed states $G_{t-1}, G_{t-2}, G_{t-T}$. The entropy of this source is given by formula (2), and shall be denoted by U bits / symbol s_{ji}, for short. P_i is then able to so encode the s_{ji} as to store $c_i^* / U - \epsilon$ symbols, where ϵ is arbitrarily small, on the average; but more than c_i^* / U symbols cannot be stored. c_i^* is a modified value for the capacity of P_i; it is c_i minus the number of bits required to store the v_i and the frequencies themselves, as well as any orders and standard comparison symbols that are necessary for the determination of U and the actual procedure of optimum encoding, such as the Shannon-Fano code.[5]

It should be noted that c_i does not change with time, and may be considered as one of the fundamental parameters which characterize P_i. c_i, then, determines the maximum number of other members with whom a given P_i may be in contact in the sense of there existing a set S_{ji}; it also determines the maximum amount of information P_i can theoretically gather about the behaviour of another member, in the sense that it determines the limit on m, the number of elements into which S_{ji} is partitioned, beyond which further partitioning is valueless. Considerations quite similar to the above can be applied to the rate r_i, which is another important parameter characterizing P_i. r_i may be regarded as the channel capacity in the above formulation, and Shannon's fundamental theorem applies there also. c_i and r_i should, of course, be related, for a large value of r_i and a small value of c_i is quite a useless combination.

Formation of Patterns.

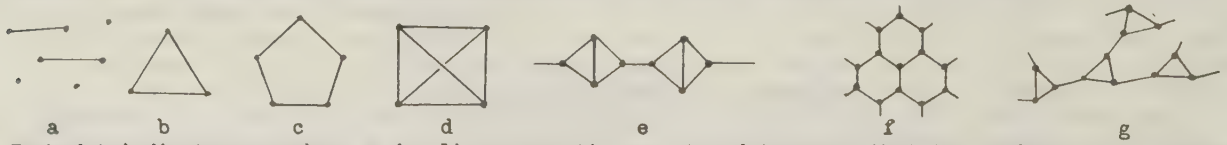
Two kinds of patterns may be distinguished: the sequence of states of the system, as observed by P_i for the last T time units, may exhibit temporal regularities; for instance, certain states may recur with a definite period. Since this kind of pattern is primarily an evolutionary phenomenon, it will be mentioned here only in relation to the second kind of pattern: the networks of members with their associated orders and efficiencies, as known to the outside observer into which the system P decomposes, may have certain topological regularities; for instance, a system of 10 members, each of whom communicates with exactly one other member, decomposes into 5 couples which act independently of one another.

To make the notions of efficiency and directed communication more precise, it is convenient to define:

$$e_{ji,t} = 1 - \frac{U_{ji,t}}{\log_2 m_{ji,t}}, \qquad e_{i,t} = \frac{1}{N}\sum_{j=1}^{N} e_{ji,t}, \qquad e_t = \frac{1}{N}\sum_{i=1}^{N} e_{i,t}$$

$U_{ji,t}$ is P_i's uncertainty about P_j, measured up to time t, according to formula (2). $m_{ji,t}$ is the number of elements in S_{ji} by time t. $e_{ji,t}$ can be taken as a measure of the efficiency with which P_j can communicate to P_i by time t. If this quantity is near 1, P_j communicates to P_i to a large extent; P_i will be quite certain as to which element of S_{ji} P_j selects. If $e_{ij,t}$ is also close to 1, then P_i and P_j are in communication with one another to a large extent. The extent of communication in such a pair may be measured by the average of the two efficiencies, e.g. $1/2(e_{ij,t} + e_{ji,t})$. $e_{i,t}$ represents the average efficiency with which P_i's environment (the system P minus P_i) communicates to P_i. e_t is an indication of the extent of communication or efficiency of the system as a whole.

Because of the limited storage capacity of P_i, the number of others with whom P_i can communicate to a specified extent is a determined quantity. As an example, suppose that each P_i can communicate with as many as 3 others. Then the following possible network patterns may be present:



a b c d e f g

Each dot indicates a member, and a line connecting any two dots means that two members are in two-way communication to a non-negligible extent. b and c are both examples of closed simple chains, where each P_i may communicate with two others; open chains of this sort may also exist, the members at the ends being of group a. d is illustrative of any polygon in which three lines emanate from each point. Patterns of type e may again appear in open or closed chains. A large number of patterns like g can easily be visualized. A system composed of a large number of members may decompose into any number and variety of such patterns; the patterns of the closed type, like a,c,d are not in communication with one another, in the sense that no individual of one communicates appreciably with some individual of the other. People in a ballroom are an illustration of this. In a pattern like e, on the other hand,

the decomposition may be regarded as one into triangular clusters which are weakly connected; one member of each cluster communicates with one member of some other one. The preceding considerations could also be applied to patterns in which the lines connecting two points are directed line segments, indicating the extent of one-way communication in the direction of the arrow, and proportional to its length. A diagram quite similar to an ordinary sociogram is obtained. It is quite clear that neither one nor two-way communication need be transitive relations, in the sense that if P_i communicates to P_j and P_j to P_k then P_i would have to communicate to P_k also. If transitivity were postulated, isolated clusters of the type a,b,c would be obtained.

In addition to the static patterns discussed above, it is possible to describe patterns of choices within a particular state of the system. If the system stabilizes, i.e. reaches a steady state, such a pattern is quite likely to obtain. In addition, a pattern may change with time, and recur, as in the case of oscillating systems; thus, a system may be descibed in terms of the communication and choice patterns at any one period, and also in terms of temporal patterns, which are essentially recurrences of the communication and choice patterns in definite periods. There is yet much to be done in this area in the direction of relating the possible decompositions of a system into patterns to the storage capacities, the value functions, and the channel capacities of the P_i.

Results

Uniform Convergence of Frequencies.

Definition: Abbreviate $f_{ji,t}(s_{ji}/\mathcal{G}_{t-1},\mathcal{G}_{t-2},\ldots,\mathcal{G}_{t-\tau})$ by $f_t(j)$. $f_t(j)$ is uniformly convergent in t, if for any rational $\epsilon > 0$, there exists a T', dependent on ϵ, but independent of j, such that $t > T'$ implies that $|f_t(j) - f_{t+\tau}(j)| < \epsilon$, for any positive τ.

This definition is evidently not satisfied by any finite sequence such as is available to the P_i unless some assumptions about the infinite sequence are inferred from the finite beginning. This is tantamount to an assumption about the regularity of P_i's behaviour, but one that is subject to continual revision and verification as the observation time progresses. A small number, $\epsilon = \epsilon_o$, could be chosen, and the above definition could be weakened to hold for all but a small fraction α of values of t which lie between T' and the largest t for which data is available.

Theorem: If $f_t(j)$ converges uniformly to $f(j)$, then $U_t = -\sum_{j=1}^{m} f_t(j) \log f_t(j)$ $(s_{ji} \in S_{ji})$ converges uniformly to $U = -\sum_{s_{ji} \in S_{ji}} f(j) \log f(j)$.

Proof: It is necessary to show that for any $\epsilon > 0$, there exists an integer $T(\epsilon)$, such that $t > T$ implies that

$$|f_t(j) \log f_t(j) - f_{t+\tau}(j) \log f_{t+\tau}(j)| = \left| \log \frac{f_t(j)^{f_t(j)}}{f_{t+\tau}(j)^{f_{t+\tau}(j)}} \right| < \epsilon . \qquad (4)$$

When $f_t(j) = 0$, define $f_t(j)^{f_t(j)} = 0$.
Let $\delta = \min_j f_t(j)$.

Choose ϵ' such that $\epsilon' < \delta\epsilon$, where ϵ is arbitrary. It is possible, by the hypothesis, to find an integer $T(\epsilon')$ such that $t > T(\epsilon')$, τ positive, imply that $|f_t(j) - f_{t+\tau}(j)| < \epsilon'$.

Since $f_t(j) \leq f_t(j)^{f_t(j)} \leq 1$,

Choosing $T(\epsilon) > T(\epsilon')$, $\left| f_t^{f_t} - f_{t+\tau}^{f_{t+\tau}} \right| \leq \epsilon'$ and $\left| \frac{f_t^{f_t}}{f_{t+\tau}^{f_{t+\tau}}} - 1 \right| < \frac{\epsilon'}{\delta}$

so that (4) is verified. Because of the uniform convergence, the summations may now be performed term by term, and the result follows.

The theorem can be generalized to the case where S_{ji} is a measurable set of finite total measure, μ, $f_t(s)$ a measurable function on S_{ji}, and

$$U_t = -\int_{s \in S_{ji}} f_t(s) \log f_t(s) \, d\mu(s)$$

The integral is an abstract Lebesgue integral. The proof is essentially the same, and based on the fact that a uniformly convergent series may be integrated term by term. Furthermore, the hypothesis can be weakened by dropping the assumption of uniform convergence, because the Lebesgue convergence theorem[7] guarantees that the operations of limit and integral may be interchanged, $|f_t(s) \log f_t(s)|$ being bounded by 1 for all s in S_{ji}.

It is of some interest to study the rate of convergence. Let $T_o(\delta\epsilon)$ be the smallest integer such that $t > T_o(\delta\epsilon)$ implies that $|f_t(j) - f_{t+\tau}(j)| < \delta\epsilon$. Then this integer also represents the smallest number of terms in the sequence $\{f_t(j)\}$, which will make (4) valid for $t > T_o(\delta\epsilon)$. This is true because if $\epsilon' < \delta\epsilon$ were chosen, $T_o(\epsilon')$, the smallest number such that $t > T_o(\epsilon')$ implies that $|f_t(j) - f_{t+\tau}(j)| < \epsilon'$, then would be larger than $T_o(\delta\epsilon)$, $T_o(\epsilon)$ being a monotonically non-increasing

function of ε .

Theorem: If $T_o(\varepsilon)$ is the step-function of ε defined above, then $t > T_o(\varepsilon)$ implies that

$$\left| \frac{U_t - U_{t+\tau}}{\tau} \right| < \varepsilon$$

Proof: Define $\Delta U_t = U_{t+1} - U_t$. Then $t' > t$ implies that $\Delta U_{t'} \leq \Delta U_t$.
To prove the above statement, suppose the contrary: $\Delta U_{t'} > \Delta U_t$. Let ε be $|\Delta U_t|$ and find $T_o(\varepsilon)$, the smallest integer such that $t_1, t_2 > T_o$ imply that $|U_{t_1} - U_{t_2}| < \varepsilon$
Clearly, $t \geq T_o$, whence $t' > T_o$. Taking $t_1 = t'$, and $t_2 = t' + 1$,

$$|\Delta U_{t'}| < \varepsilon \quad \text{or} \quad |\Delta U_{t'}| < |\Delta U_t|, \text{ which is a contradiction.}$$

Now, $|U_t - U_{t+\tau}| \leq |U_t| \tau$.
For $t > T_o$, $|\Delta U_t| < \varepsilon$ and the result follows.

By means of this theorem, the smallest number of terms can be found which stabilize the uncertainty in the sense that its time rate of change is less than an arbitrarily prescribed ε . P_i can further reduce his uncertainty about P_j by repartitioning S_{ji} into a finer net.

If, in fact $f_{ji,t}(s_{ji}/\mathcal{G})$ is a uniform distribution, and S_{ji} is, for example,

$$S_{ji} = \left\{ \left[0, \tfrac{1}{m}\right), \left[\tfrac{1}{m}, \tfrac{2}{m}\right), \ldots \left[\tfrac{m-1}{m}, 1\right] \right\}$$

then it is easily shown that the best repartition, in the sense of removing the most uncertainty is:

$$\left\{ \left[0, \tfrac{1}{2m}\right), \left[\tfrac{1}{2m}, \tfrac{2}{2m}\right), \ldots \left[\tfrac{2m-1}{2m}, 1\right] \right\}$$

For, $U_{ji,t} = \log m$; letting the new $m' = f(m)$, $\Delta U_{ji,t} = \log f(m) - \log m$. To obtain $f(m)$ such that $\Delta U_{ji,t}$ is maximum,

$$\frac{d\Delta U_{ji,t}}{dm} = \frac{1}{f(m)} \cdot \frac{df(m)}{dm} - \frac{1}{m} = 0$$

whence $\log f(m) = \log m + \log 2$, since $f(1) = 2$. Therefore, $f(m) = 2m$.

Two - member systems.

Are there any relations which govern the uncertainties which follow from the assumption that each P_i tries to maximize the value to him? It is instructive to consider the special case of $N = 2$. Assume that the members, herein also called players, have played so long that P_1 knows that

if he chooses	$S_{11}^{(1)}$	and P_2 chooses	$S_{21}^{(1)}$	, he wins	$V_1^{(11)}$;
"	$S_{11}^{(1)}$	"	$S_{21}^{(2)}$	"	$V_1^{(12)}$;
"	$S_{11}^{(2)}$	"	$S_{21}^{(1)}$	"	$V_1^{(21)}$;
"	$S_{11}^{(2)}$	"	$S_{21}^{(2)}$	"	$V_1^{(22)}$.

Assume also that P_2 knows that

if he chooses	$S_{22}^{(1)}$	and P_1 chooses	$S_{12}^{(1)}$	, he, P_2, wins	$V_2^{(11)}$;
"	$S_{22}^{(1)}$	"	$S_{12}^{(2)}$	"	$V_2^{(12)}$;
"	$S_{22}^{(2)}$	"	$S_{12}^{(1)}$	"	$V_2^{(21)}$;
"	$S_{22}^{(2)}$	"	$S_{12}^{(2)}$	"	$V_2^{(22)}$.

Suppose further that P_1 "guesses correctly" with frequency q_o that P_2 chooses $S_{21}^{(1)}$, hence $S_{21}^{(2)}$ with frequency $1 - q_o$. It would be reasonable to assign to q_o some a priori value such as

$$q_o = \frac{Max\left[(V_2^{(12)} + V_2^{(22)}), (V_2^{(11)} + V_2^{(21)})\right]}{Max\left[V_2^{(11)}, V_2^{(12)}, V_2^{(21)}, V_2^{(22)}\right]}$$

To simplify the notation, as well as the calculations, let it be further assumed that $s_{21} = s_{11}$, $s_{12} = s_{22}$ and $v_2 = -v_1$. The problem now assumes the form of a simple rectangular two-person, zero-sum game, with a 2 x 2 utility matrix, in which each player does not choose in complete ignorance of the other's choice. It is not difficult to see how the results can be extended.

Let p_o be the frequency with which P_1 chooses s_{11} such that his expectation,

$$E_1 = p_o\left[V_1^{(11)} q_o + V_1^{(12)}(1-q_o)\right] + (1-p_o)\left[V_1^{(21)} q_o + V_1^{(22)}(1-q_o)\right]$$

is maximized, p_o being a function of q_o, $f(q_o)$. If $V_1^{(12)} > V_1^{(22)}$, the solution of the resulting differential

equation on f, subject to the initial condition $f(0) = 1$, is $p_o = Aq_o + B$, where A and B are functions of the v's. If $V_1^{(11)} < V_1^{(12)}$, the trivial result $p_o = q_o$ obtains.

If, now P_2, after some time, chooses so as to alter P_1's frequency of "guessing correctly" from q_o to q_1, q_1 is a function of p_o such that P_2's expectation is maximum. This statement is now extended to an inductive statement, relating p_t to q_t and q_t to p_{t-1}, expressible as the following pair of recursion relations:

$$p_t = \frac{V^{(11)} - V^{(12)}}{V^{(21)} - V^{(22)}} q_t + 1 \qquad q_t = \frac{V^{(11)} - V^{(21)}}{V^{(12)} - V^{(22)}} p_{t-1} + 1$$

The solution to these, in terms of the initial value q_o, or p_o is given by:

$$p_t = K^t p_o + C \sum_{k=0}^{t-1} K^k = \frac{C + (p_o - C)K^t - p_o K^{t+1}}{1 - K} \tag{5}$$

where

$$C = \frac{V^{(11)} - V^{(12)} + V^{(21)} - V^{(22)}}{V^{(21)} - V^{(22)}} \qquad K = \frac{(V^{(11)} - V^{(12)})(V^{(11)} - V^{(21)})}{(V^{(21)} - V^{(22)})(V^{(12)} - V^{(22)})}$$

There are certain restrictions of the v's to insure that p_t remains between 0 and 1.

These are $0 \leq C \leq 1 - K$, or in terms of the v's:

$$0 \leq (V^{(11)} - V^{(12)} + V^{(21)} - V^{(22)})(V^{(12)} - V^{(22)}) \leq (V^{(21)} - V^{(22)})(V^{(12)} - V^{(22)}) - (V^{(11)} - V^{(12)})(V^{(11)} - V^{(21)})$$

If the utility matrix is symmetric, $V^{(21)} = V^{(12)}$, and $K = \left(\frac{V^{(11)} - V^{(12)}}{V^{(12)} - V^{(21)}}\right)^2 = 1$, $C = 0$, and $p_t = p_o$ for all t. On the other hand, if $|K| < 1$, the dependence of p_t on the initial value disappears. Under very mild restrictions on the v's, the sequence $\{p_t\}$ converges uniformly, and the previous results apply. Thus, if $|K| < 1$, a limiting expression for the entropy can be found, which is independent of the a priori uncertainty of P_2 about P_1, or P_1 about P_2.

$$U = \frac{C}{1-K} \log \frac{1-K}{C} + \frac{1-K-C}{1-K} \log \frac{1-K}{1-K-C}$$

A similar relationship holds for the other player, with q_t in place of p_t. When the restrictive assumptions which were made above, and the general case of N instead of 2 players is considered, a system of N simultaneous difference equations must be solved. The results will then permit the computation of the efficiencies, and the extent of communication, e_t, in terms of the assumed value functions for each P_i. The capacities c_i and r_i will then determine the possible patterns of communication, as discussed before, and permit the computation of the efficiencies associated with each.

Conclusions.

The main limitations of this model are that they represent over-simplifications of most existing organizations, but these can be gradually dropped as the theory is further developed. The manner in which the outside observer determines the value functions, storage capacities, is similar to the way in which psychologists test subjects to obtain data about subjective judgments, rates of learning, capacities of retention, etc. The limiting processes which are used in this model are subject to the same criticism as those in von Mises' frequency approach to probability.

The chief values of this model may prove to be the readiness with which the methods involved are adapted to treatment by digital computers, particularly with regard to the data matrices and the difference equations. The memory requirements will, of course be large, to handle any organizations of interest. It also becomes easy to formulate many problems, such as the analysis and synthesis or organizational structures, group coherence, leadership, learning, etc., which could not be simply formulated otherwise, and experiments to test the model in such applications are readily suggested and designed.

Bibliographical References

(1) Arrow, K., Social Choice and Individual Values, Wiley, N.Y. 1950
(2) von Neumann, J., & Morgenstern, O., Theory of Games and Economic Behaviour, Princeton, 1947
(3) Bush, R., & Mosteller, F., "A Stochastic Model With Applications to Learning", Ann, Math. Stat. 24,4
(4) Shannon, C., & Weaver, W., The Mathematical Theory of Communication, Univ. Illinois, Urbana, 1949
(5) Landau, E., Foundations of Analysis, Chelsea, N.Y. 1951
(6) Fano, R.M. "The Transmission of Information", Tech. Report No. 65, Res. Lab. of Electronics, M.I.T.
(7) Halmos, P., Measure Theory, Van Nostrand Co., N.Y. 1950

SIMULATION OF SELF-ORGANIZING SYSTEMS BY DIGITAL COMPUTER*

B. G. Farley and W. A. Clark
Lincoln Laboratory, Massachusetts Institute of Technology
Cambridge, Massachusetts

ABSTRACT

A general discussion of ideas and definitions relating to self-organizing systems and their synthesis is given, together with remarks concerning their simulation by digital computer. Synthesis and simulation of an actual system is then described. This system, initially randomly organized within wide limits, organizes itself to perform a simple prescribed task.

INTRODUCTION

Information systems whose response to a given class of inputs changes with time in accordance with specified criteria which are chosen to correspond roughly to the "self-organizing" concept have been the subject of considerable interest.[7,13] Several mechanisms have been constructed or described which are, "self-organizing" to some extent,[1,6,8,11] and some work has been published on computer-programmed learning, such as that by Oettinger.[5] Recently, McKay has communicated ideas related to some of those to be discussed here.[4]

The work to be described was undertaken in an attempt to clarify certain ideas related to such systems, and to try to gain some insight into their synthesis by simulation of specific systems using a digital computer. Although the work is in an early stage, it is believed that results so far have exhibited some very interesting properties of a particular system, and have demonstrated the usefulness of computer simulation methods in studies of this kind where systems are likely to be so complex that analytical solutions are difficult or impossible, or do not furnish much information until leads are suggested by actual experience.

The work will be presented in three parts. First a general discussion will be given in which definitions will be made. Second, the definitions will be applied to an experimental system. Third, the details of a self-organizing system and a description of computer techniques used in its simulation will be given.

General Considerations and Definitions

In order to make our ideas and definitions precise, and at the same time as general as possible, it is convenient to introduce a mathematical framework to aid in discussion.

We will deal first with a general system as shown in Figure 1. Inputs p_j from the left are transformed into outputs q_j on the right. As indicated in Figure 1, both input and output lines may be multiple. In what follows, the symbols p_j and q_j will refer to specific, complete configurations on these multiple lines, finite or infinite in time. No loss of generality will result if all signals are reduced to a binary equivalent. As an example, then, if there are three input lines, a certain input might be defined as

$$p_3 = \begin{cases} 0110001011001 \\ 1001101011001 \\ 0110101000000 \end{cases} \qquad (1)$$

time increasing to the right. Such a configuration will be called a time-channel pattern.

The transformation T will be allowed to change with time, and we are interested in this change in so far as it exhibits organizing properties with respect to T. To define properties of this type we will fix our attention on a particular class C of inputs p_j, $j=1,2,\ldots,n$ and their corresponding outputs q_j. Each member of this class will usually be finite in length.

We may then break the transformation T down into a class of transforms

$$T = \{T_1, T_2, \ldots, T_n\} \qquad (2)$$

*The research in this document was supported jointly by the Army, Navy, and Air Force under contract with the Massachusetts Institute of Technology.

where the set of equations

$$T_1 p_1 = q_1$$
$$\cdots \qquad (3)$$
$$T_n p_n = q_n$$

serves to define $T_1 \cdots T_n$ If the system contains sources of noise or produces spontaneous outputs, the transformations $T_1 \cdots T_n$ will be defined statistically as averages over an ensemble of identical systems started in the same initial state.

Now, in order to discuss one or more properties of such a system dependent on time, it is only necessary to choose a measure m specifying the properties in question and apply it to T at succeeding times. In many cases, the measure m(T) will of course depend upon only one, or a few, of the T_j's.

We will consider that the instrument of time change of T is within the system. As an aid to visualization, Fig. 2 shows the system broken down into two components, one of which contributes primarily to the transformation T itself, and the other, called the modifier, has the primary function of producing changes in T. The double line between the modifier and T represents the agency of the modification, while single lines show information paths. While a sharp dichotomy of function between T and modifier has been indicated, it is not intended to exclude systems in which the modifier contributes to the transformation or modifies itself. We will consider everything outside the dashed lines as "environment," although it should be noted that the exact path of such boundaries is arbitrary.

It may be of some interest to suggest as an illustrative example how the above model as described might be used to describe situations which approximate psychological definitions of "learning."

Learning behavior by an organism may be defined for the purposes of psychology as a positive change in the proficiency of performance of one or more tasks as a result of prescribed experience.[9]

In terms of our model, we may describe this as follows. A number of input patterns are chosen, and presented to the organism in prescribed orders and times to provide the required "experience." One or more of these inputs are designated as performance tests or tasks, and a suitable proficiency measure, such as a test score, is constructed. This score corresponds to the measure we have attached to the general transformation. If the measure increases as a result of presentation of "experience" inputs, and does not increase otherwise, the organism is said to learn.

The provision that the measure should increase only as a result of presentation of "experience" rules out as learning systems those in which the modifier operates to increase a measure without information inputs. Control experiments may of course be required to rule out such cases.

It should be noted that learning thus defined is relative to the input class and measure chosen, and the experience prescribed. By varying these parameters, various kinds of learning may be defined. For example, transfer learning requires altered experiences or measures of new performance (perhaps with special control); learning with relatively short performance inputs is called conditioning, while that with long performance inputs is called serial learning. More precise definitions would require close examination of the variable parameters. This task is complicated for the psychologist by the fact that he is dealing with organisms no two of which are alike.

Some competing theoretical interpretations of learning may also be referred to the model. For example, reinforcement and non-reinforcement theories make different assumptions as to the nature of the modifier organization. "Perception" theorists make use of "perceptual systems or fields" which are not explicitly represented in our model as presented here.[9]

No matter how complex the organization of a system such as we have been discussing, it can always be simulated as closely as desired by a digital computer as long as its rules of organization are known. This possibility is indicated for example, by the work of Turing.[10] This means that the action of any system can be studied even though it is too complex for mathematical analysis. Furthermore, the computer offers unparalleled flexibility in such work, since any part of a simulated system may be quickly and easily modified to judge the effect of the change. There is of course the disadvantage that present computer simulation takes pace serially in time, so that even with very fast computers considerable time may be required to simulate highly complex systems. Balanced against this disadvantage, however, is the fact that the initial programming for simulation in general requires a great deal less time than actual construction of an analogue device even if this is feasible, so that for a very wide class of problems the net advantage in both time and cost lies on the side of a computer simulation method, and for an additional large class this method is the only feasible one, at least until the system is reasonably well understood.

The work to be described was undertaken partly to examine the problems encountered in such simulation. Furthermore, it was desired to answer two questions: (1) Can a transformation, initially organized at random between rather wide limits, be provided with a modifier which will cause it to become organized, as a result of experience, to perform a prescribed task? (2) Can such a system be generalized to organize itself to perform any of a rather wide class of tasks? The work to be described is still in an early stage, but has resulted in the synthesis of a system which it is believed fulfills the requirements of the first question.

Application to an Experimental System

In seeking to synthesize systems along the lines discussed above, it is natural first to choose a transformation with promising transforming and modifiability possibilities and then try to discover suitable modifiers. Preliminary investigation showed that transformations composed of interconnected active non-linear elements with definite thresholds as indicated in Fig. 3 have interesting transforming properties. For example, such a net of elements can change a time-channel pattern into a space pattern of active elements, and if it is complex enough, can do this uniquely for a given class of patterns. Furthermore, enough variable parameters are available in the net to give it useful modifiability properties. Networks resembling those under discussion exist naturally, and have a great intrinsic interest, namely networks of nerve fibers or neurons.[2] It was therefore decided to use non-linear elements possessing many of the known properties of neuron nets as an experimental transformation.[3] The details of the net and associated modifier will be presented later, but first the simple task chosen for performance, the measure of proficiency used, and the prescribed experience, will be described in terms of the framework already discussed.

First a randomly connected net is arbitrarily divided into four groups of elements designated as groups I_a, I_b, $O(+)$, and $O(-)$. These symbols stand for input groups "a", "b", and output groups "+", and "−", respectively.

Two input patterns, p_1 and p_2 are considered. The first, p_1, may be represented by the following scheme,

$$p_1 = \begin{cases} \left.\begin{matrix} \ldots00100100100\ldots \\ \ldots00100100100\ldots \\ \ldots\ldots\ldots\ldots\ldots \end{matrix}\right\} I_a \\ \left.\begin{matrix} \ldots00000000000\ldots \\ \ldots00000000000\ldots \\ \ldots\ldots\ldots\ldots\ldots \end{matrix}\right\} I_b \end{cases} \tag{4}$$

which indicates that the same periodic input is applied to every element of I_a, and that no input is applied to I_b. The input p_2 is identical except that the roles of I_a and I_b are reversed. When p_1 is applied, the transformation called T_1 is active, and T_2 is active when p_2 is applied, in accord with equation (3). In order to define a proficiency measure, we proceed as follows: Let $n(+)$ be the number of elements active during a given time interval in group $O(+)$, and $n(-)$ the number active during the same interval in group $O(-)$.

The measure $m(T)$ will be composed of two components, m_1 and m_2.

$$m(T) = \{m_1, m_2\} \tag{5}$$

where

$$m_1 \equiv m_1(T_1) = \overline{n(+) - n(-)} \tag{6}$$

$$m_2 \equiv m_2(T_2) = \overline{n(-) - n(+)}$$

and the bar denotes a time average over a fixed interval.

Note that m_1 is defined above only when T_1 is active, and similarly for m_2 and T_2. Organization will be said to occur if both m_1 and m_2 increase.

In other words, we may consider an output formed by the accumulated difference of the numbers of cells active in $O(+)$ and $O(-)$. Presentation of experience will be externally arranged so that p_1 is applied whenever the output is positive, ($O(+)$ predominates) and p_2 whenever the output is negative, ($O(-)$ predominates). If the output remains near zero for a specified length of time, it is externally "forced" from zero by adding to the output difference in alternately positive and negative directions. Thus the whole mechanism is similar in some respects to a servo which must learn to return to zero when displaced, training experience being given alternately on either side of zero, and increasing organization being manifested by an increasing rate of return. The patterns p_1 and p_2 provided by the environment may be said to enable the mechanism to "sense" the position of its output.

The modifier which causes the measures m_1 and m_2 to increase was determined largely empirically. It operates on various parameters of the net in a way to be described later. Information for the operations

of the modifier is generated internally in this simple case in a manner which essentially computes m_1 and m_2. However, it should be mentioned that in the general case this may not necessarily be true. That is, the modifier may use information related to the organization measure, but computed in some entirely different manner.

Details of Experimental System and Simulation Program

The general properties of the particular transformation and associated modifier with which the initial simulation work has dealt have been presented. This description will now be expanded and related to the computer simulation techniques which were developed for the Memory Test Computer of the Lincoln Laboratory of M.I.T. A note on the characteristics of the computer may be of general interest: MTC is a 16 bit, parallel machine with a coincident current magnetic-core memory of 4096 words and an operating speed of about 90,000 single-address add-type instructions per second. Its principal input device is a Ferranti Photoelectric Reader for punched paper tape and output equipment includes a standard flexowriter and several cathode ray tubes for displays which may be photographed.

The transformation system has been described as a network of non-linear elements in which the pathways or connections are randomly established. In this and other parts of the program random processes play an important part and should be discussed in more detail. MTC does not have access to a random element, but there exist many accredited computation routines which generate number sequences in which the values of the terms are distributed in a nearly statistically homogeneous manner. The pseudo-random number generator routine which was used develops the n^{th} terms, R_n, by means of the recursion relation

$$R_n = R_{n-1} + R_{n-k} \quad \text{(Sum modulo p)} \tag{7}$$

The series initially is "primed" with k terms chosen from a table of random numbers.

To connect network elements at random a matrix P_{ij} expressing the probability that i connects j is established for the class of networks under consideration. In the systems to be discussed, the simple case $P_{ij} = K$, constant for all i, j was chosen, but more generally the connection probability might depend on i, j or any particular characteristics of network elements i and j. For each pair of network elements a pseudo-random number in the interval ab is then generated and a test is made to determine whether the number lies also in a subinterval ar of ab where r is so chosen that the ratio of (r-a) to (b-a) is the probability P_{ij}. Since the pseudo-random numbers are uniformly distributed in the interval, this test yields positive results with a mean relative frequency equal to the required probability. For each positive test result, a connection is established and in this way a specific connection matrix, ($c_{ij} = 1$ if i connects j, 0 otherwise), is set up for the given network. K will be called the connectivity of the net.

With each connection there is associated a sixteen-state weight, w_{ij} which determines the excitation value on j of activity transmitted from i via this connection (see fig. 3). These weights may in general be drawn from a distribution in the manner discussed above, although in the example presented later these weights were chosen equal and set initially at mid-value.

For each element in the network, one row of the connection matrix (representing pathways from the element) and a list of associated weights are stored in the computer memory. This requires breaking the matrix row into 16-bit words and also packing four of the 4-bit weights into one word for storage economy, and much of the computing time is consumed in the unscrambling and repacking of these words during the simulated operation of the network.

Similarly, with each network element there is stored a list of characteristics such as threshold, time constants, etc. selected from appropriate distribution functions, and addresses and counters required by the simulation process. These quantities occupy another six 16-bit words. The total storage requirement, however, is determined largely by the connection matrix and associated weights, since for these the required capacity increases with the square of the number of network elements. The 4096 registers of the MTC memory limit the size to a network of about 128 elements with connectivity of 0.4. The time required to generate such a net is approximately ten seconds, or about 900,000 operations. The complete simulation program occupies about 1500 registers; the remainder of the storage is occupied by the characteristics of the network.

In order to elaborate on the characteristics of the network elements it is necessary to discuss more completely the transient behavior of the element during excitation. This transient state of activity occurs whenever the excitation level exceeds the threshold of the element. After a small time delay, the element transmits by simultaneously increasing the excitation level of all other elements to which it is connected as indicated by its associated row in the connection matrix. At the beginning of this delay interval, the threshold rises to a value which is large enough to prevent a second activation during the interval. At the end of this interval, suggested by the refractory period in neurons, the element recovers sensitivity as it threshold decays exponentially to a minimum value characteristic of the element,

measured relative to an adjustable bias level for the network as a whole. The threshold function, $h_j(t)$, for the j^{th} element may thus be represented as effectively infinite during the refractory interval and

$$h_j(t) = h_{max}\exp(-a_j t) + h_{jmin} + h_{bias}(t) \qquad (9)$$

otherwise, where a_j is the threshold decay constant. The comparison of excitation with threshold occurs in the presence of gaussian noise such that a high level of excitation increases the probability that an element will "fire" but will not in general completely determine the instant of firing. The behavior of the network becomes completely determinate as the mean-square amplitude of the noise, which, like the bias level, is controlled by the modifying sub-system, is reduced to zero. The gaussian distribution is approximated, as suggested by the central limit theorem, by averaging a set of four pseudo-random terms for each term of the "gaussian" set.

When several elements simultaneously transmit to the same element, the change in excitation of the affected element is chosen to be the sum of the weights of the active connections, although a more complicated function of the weights might be used. In addition, the total excitation level at the affected element decays exponentially with a time-constant characteristic of the element. Thus, activity pulses arriving within a small time-interval of one another partially combine in excitation value in a manner related to the observed temporal summation effects in neurons. The change in excitation, $\Delta s_j(t)$, at the j^{th} element at time t may then be written

$$\Delta s_j(t) = -b_j\, s_j(t-1) + \sum_i w_{ij} \qquad (9)$$

where the summation extends only over elements which transmitted at t-1 excluding, in the model chosen, i=j, and b_j is the excitation decay constant.

The characteristics stored in the computer memory for the j^{th} network element can now be enumerated in summary:

(1) Type of element, i.e. number of the group I_a, O(-) etc. to which the element is assigned.

(2) Time delay, which determines the refractory period, and also, in the simple model chosen, the delay between firing and transmitting (equal for all pathways from the transmitting element).

(3) Minimum threshold, h_{jmin}

(4) Threshold decay constant, a_j

(5) Excitation decay constant, b_j

(6) Connection Matrix row, c_{jk}, k=1,2,..,n where n is the number of elements in the network.

(7) Those connection weights, w_{jk}, for which $c_{jk}=1$

It should be pointed out that as yet there has been no systematic evaluation of the effects of varying thresholds, decay constants, and time delays. Their inclusion in the set of characteristics does, however, illustrate the degree of complexity of the model being simulated.

In the simulation program the time variable is quantized into equal intervals of about one-eighth of a refractory period. This time parameter is, in effect, frozen until the program has scanned through storage, calculating values of threshold, excitation, etc. for each element, after which it is advanced to the next larger value. The real time consumed per "time" interval in carrying out these calculations for a net of 128 elements with connectivity of 0.4 is about one second, varying from interval to interval according to the amount of activity within the network.

Activity is introduced into the network by increasing by a large fixed value the excitation level of those input elements, and at those times, indicated by the presence of "ones" in an input pattern similar to the p_1 of eq. (4). The output, as described earlier, is formed simply by counting the number of transmitting elements in the output groups O(+) and O(-) during each time interval. The difference of these numbers, $n_t(+) - n_t(-)$, defines the changes in the output N_t of the net so that

$$N_{t+1} = N_t + n_t(t) - n_t(-) \qquad (10)$$

The computer program is arranged to plot N_t against t directly on one of the display scopes and a time exposure photograph records the trace. A typical output record appears in fig. 5. In order to automatize the process of presenting input patterns, the simulated external system is so arranged that $N_t > +N'$ results in pattern p_1 and $N_t < -N'$ produces p_2 where N' is a small positive number. If N_t remains in the null interval between $-N'$ and $+N'$ for a specified period (chosen long enough to allow residual activity to attenuate) the program displaces N_t to some value $+N'' > N'$ with alternate trial displacements to $-N''$. These displacements will be seen as the discontinuities in fig. 5.

The action of the modifying system is best described by means of the flow diagram of fig. 4. If a "contributive connection" is defined as any active connection to a "fired" element which may have contributed to the firing of the element during an immediately previous fixed time, the modifier increases

the weights of contributive connections when the magnitude of the output has just decreased and decreases these weights if the magnitude of output has just increased, subject to upper and lower bounds of weight value. Note that weights are changed without regard to their individual influence on the output, and improvement in performance results from what might be termed "statistical cooperation." In addition, the modifier manipulates the threshold bias level and the noise level within the net, the former by gradually lowering bias until activity starts (principally to prevent self-sustained activity, which is difficult to control) and the latter to allow noise-initiated activity to scan, in effect, new activity modes of the network when required. Bias control of this sort may be considered use of a "field" parameter, in contrast to use of local cell parameters.

A Small Network Example

An example of an eight element network of 0.75 connectivity will now be given. To simplify the network for illustrative purposes, the elements are divided into four equal groups and numbered so that elements in the same group are represented by successive rows in the connection matrix. In this example, a_j=0.25, b_j=0.50 for all j, and the refractory delays were all equal to 2 time units.

Fig. 5 shows the history of the output of this network during the organization process requiring approximately 15 minutes of computer time. (The graph is redrawn from a set of photographs which were unsuitable for reproduction). Up to the point marked "modifier activated", the behavior of the unaltered transformation is seen to be slowly divergent for both positive and negative test displacements. Figs. 6a through 6d show the weight matrix sampled at the times indicated by points labeled "a" through "d" in fig. 5. The numbers appearing in these matrices are in octonary form and will be seen to change substantially during the organizing process.

It will be noted that the changes primarily affecting the output occur in the enclosed boxes; weights in boxes I_a O(+) and I_b O(-) tend to increase while those in boxes I_a O(-) and I_b O(+) tend to decrease. It can be seen the return to zero of the output gradually improves from a condition of divergence to increasingly rapid convergence as the matrix changes progress.

A total of perhaps 30 randomly organized nets of this type with various connectivities have actually been tried, the largest of which contained 64 elements with K=0.75. All but 3 or 4 have been organized successfully by the modifier, the failure being due to lack of essential connections or other special properties sometimes resulting from the wide variability of the random process.

It might also be of interest to note that exploratory experiments have been made to examine the effect of damage on these nets after organization. Indications are that arbitrary destruction of at least 10% of the elements may be sustained without impairment of performance.

Conclusion

We have now described a general formulation of the self-organizing concept, and a synthetic example of a system which organizes itself to perform a simple task.

Although the experimental system was composed of elements having properties similar in many respects to the known properties of neurons, it is not claimed at this stage that the results are of neurophysiological significance. However, it is believed that the results do show the great usefulness of computer simulation methods in this and other fields where systems of great complexity are encountered. Not only will simulation methods produce specific knowledge, but it is believed that they should also eventually yield enough information about given types of systems to make more general formulations possible. For example, enough experience has not yet been gained about the present experimental system to understand what features are necessary under given conditions, but it is believed such information can be elicited by an extension of the present methods.

As mentioned earlier, the gradual organization of the system to utilize the patterns p_1 and p_2 to change an output in opposite directions implies a primitive "recognition" of these patterns. It is also found experimentally that after organization other patterns also have effects like p_1 or p_2. In other words patterns are classified together by such a transformation. It is to be hoped that, using a more complex modifier, this type of behavior can also be organized and controlled, leading to systems which effect classifications and generalizations. Success in this respect should make possible systems which can organize themselves to perform in an environment presenting a rather wide variety of tasks.

Acknowledgment

The authors wish to express their appreciation to F. A. Webster for many valuable discussions, and also to those responsible for the operation of the Memory Test Computer for their very helpful cooperation.

REFERENCES

1. Ashby, W. R. Design for a Brain, Wiley, 1952
2. Brazier, M. A. B. The Electrical Activity of the Nervous System, MacMillan, 1951
3. Hebb, D. O., The Organization of Behavior, Wiley, 1949
4. McKay, D. M., (to be published) Oral communication at conference on Human Communication and Control, M.I.T., June 1954.
5. Oettinger, A. G., "Programming a Digital Computer to Learn" Phil. Mag. 43, 1243-1263
6. Shannon, C. E., "Presentation of Maze-solving Machine", Transactions of the Eighth Cybernetics Conference of the Macy Foundation, 1952, 173-180.
7. Shannon, C. E., "Computers and Automata", Proc. I.R.E., 41, 1234-1241 (Oct. 1953)
8. Shimbel, A., "Contributions to the Mathematical Biophyics of the Central Nervous System, with Special Reference to Learning," Bull. Math. Biophysics, 12, 241-274
9. Stevens, S. S. (ed.) Handbook of Experimental Psychology, Wiley, 1951, Chap. 16
10. Turing, A. M., "On Computable Numbers, with an Application to the Entscheidungspreblem" Proc. Lond. Math. Soc., 24, 230-265, (1936)
11. Walter, W. G., The Living Brain, Norton, 1953
12. Wilkes, M. V., "Can Machines Think", Proc. I.R.E., 41, 1230-1234 (Oct. 1953)

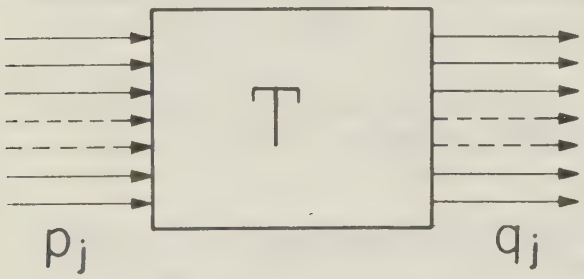


Fig. 1 - General transformation.

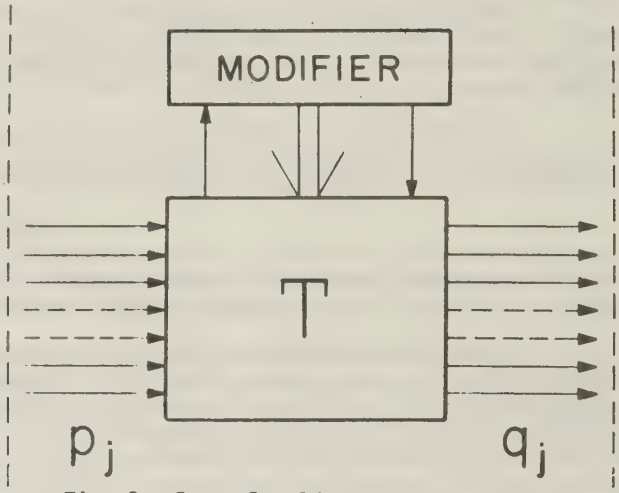


Fig. 2 - General self-organizing system.

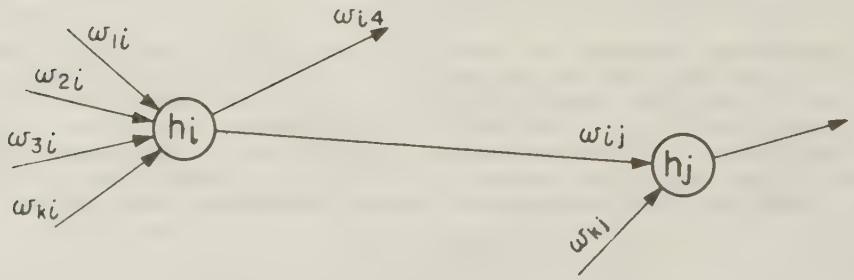


Fig. 3 - Typical section of network showing weights, w, and thresholds, h, associated with nonlinear elements i and j.

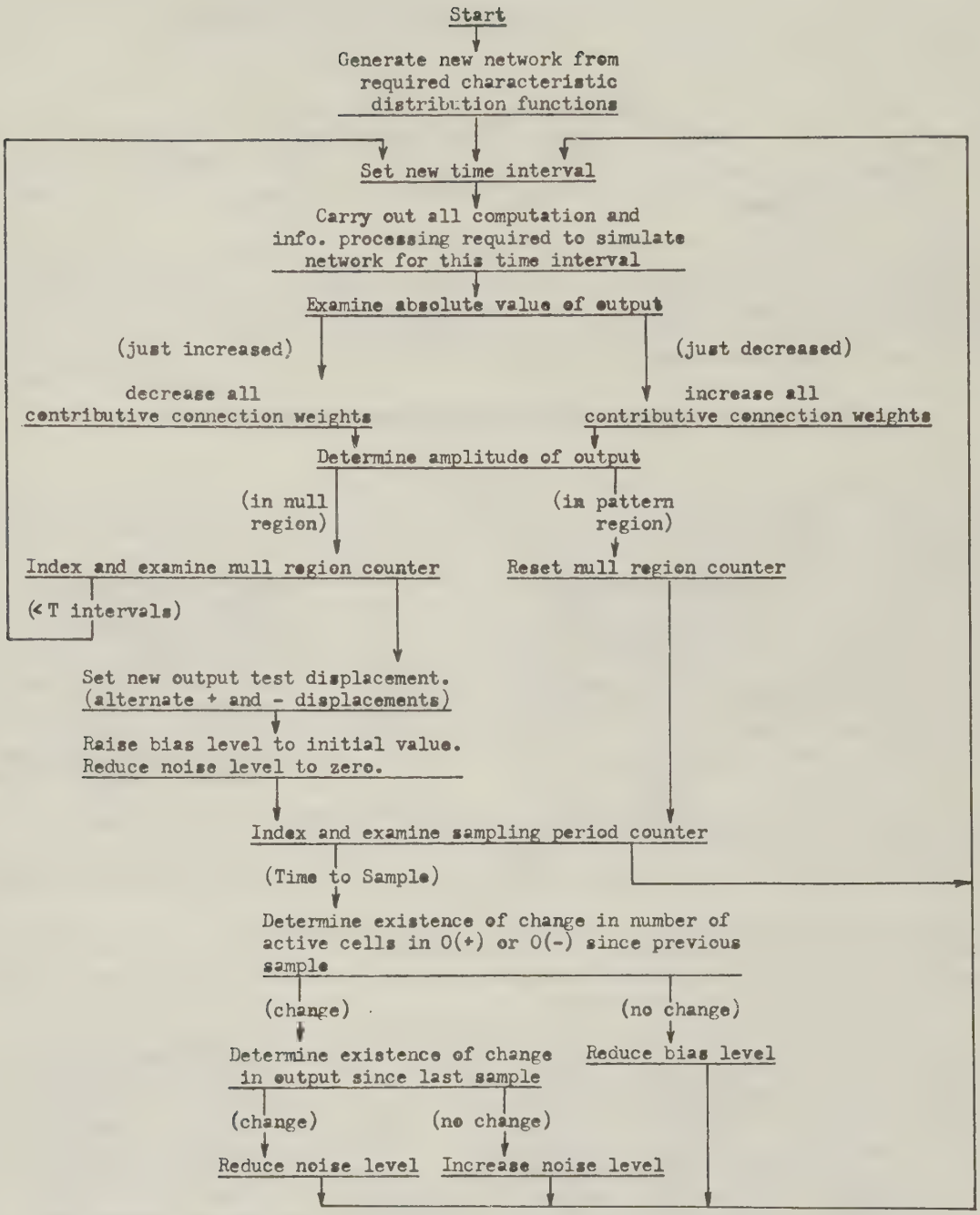


Fig. 4 - Simplified computer simulation flow diagram emphasizing modifier.

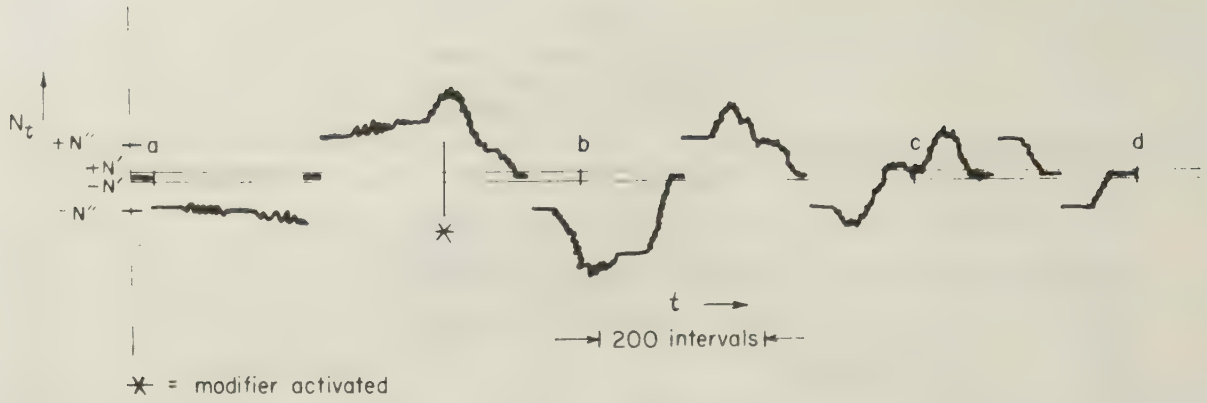


Fig. 5 - Output record of an 8-element network; K = 0.75.

(a)

	I_a	O(−)	O(+)	I_b
I_a	07 07 07	07 07 07	07 07 07	07 07 07 07
O(−)	07 07	07 07 07 07	07 07 07 07	07 07 07
O(+)	07	07 07 07	07 07 07	07 07 07 07
I_b	07 07 07	07	07 07	07 07 07

(b)

	I_a	O(−)	O(+)	I_B
I_a	07 10 07	07 11 03	04 11 02	11 11 02 07
O(−)	07 08	07 06 06 07	01 01 01 01	07 01 02
O(+)	04	14 17 10	10 01 07	12 07 02 05
I_b	10 05 12	14	06 01	07 07 06

(c)

	I_a	O(−)	O(+)	I_b
I_a	07 10 07	02 01 01	15 07 17	01 16 01 07
O(−)	16 04	07 04 02 07	11 14 07 10	07 01 10
O(+)	13	03 11 01	03 11 07	02 07 01 12
I_b	17 10 01	11	16 05	07 07 01

(d)

	I_a	O(−)	O(+)	I_b
I_a	07 03 07	02 03 01	02 03 16	01 14 01 07
O(−)	15 01	07 04 03 07	11 12 01 02	07 01 06
O(+)	15	03 12 01	04 04 07	03 07 01 11
I_b	17 17 01	16	06 02	07 07 01

Fig. 6 - The connection weight matrix for the 8-element network sampled at points labelled "a" through "d" in Fig. 5.

A STUDY OF ERGODICITY AND REDUNDANCY
BASED ON INTERSYMBOL CORRELATION OF FINITE RANGE

Satosi Watanabe
United States Naval Postgraduate School
Monterey, California

Abstract

Some of the basic concepts of information theory are critically reviewed in the light of a generalized formulation of the theory of Markoff's chains, in which the initial and final states are sequences of symbols of different lengths, and occurrence of symbols is governed by inter-symbol correlation probability of finite range. In particular, the conditions of ergodicity and the structure of "ergodic subsets" of sequences of arbitrary length are carefully discussed. A mathematical method is developed to determine the "range" and "strength" of inter-symbol correlation. A brief summary of the content is given at the end of Section 1.

Introduction

The aim of this paper is to clarify some of the basic, but often carelessly used concepts of information theory, viz., the concepts of ergodicity, intersymbol correlation and redundancy. There are two approaches to this problem-complex pertaining to probability. One is an empirical point of view, and probability here is understood in its statistical aspect. The other is an a priori point of view which deals with probability mainly in its predictive aspect. In the first standpoint, the entire population of messages in a language is supposed to be given, and the various probabilities are calculated by the actual frequencies of individual symbols or those of sequences of symbols. According to this method, a unique value of the probability of appearance of a given symbol or a given sequence can be statistically determined. In the second point of view, an ensemble of messages is supposed to be engendered by the given correlation probabilities starting from a given initial symbol or a given initial sequence of symbols. In this case, the existence of a unique, non-vanishing value of the probability of appearance of a given symbol or a given sequence is not guaranteed, for it may vanish with increasing length of messages, and it may depend on the initial condition. Thus, the problem of ergodicity acquires foremost importance in this approach.

Our section 2 dealing with the problem of ergodicity is therefore developed in the framework of the second point of view. Once the nature of the ergodicity condition is clarified and this condition is assumed to be fulfilled, then a smooth passage from the second point of view to the first becomes easy. Thus, our section 3 on redundancy can be interpreted in either point of view.

It is not implied by the foregoing paragraphs that the problem of ergodicity is irrelevant to the first standpoint or cannot be formulated in the framework of this standpoint. The situation is that the nucleus of the problem under consideration can be exhibited more directly and naturally in the second point of view.

The usual theory of Markoff's chains, which is based on transition probabilities from one state to another, is extended in this paper to the case where the probability $Q(a_1, .., a_{\nu-1} \mid a_\nu)$ of symbol a_ν appearing in a message is dependent on the $(\nu - 1)$ immediately preceding symbols, ν being the range of intersymbol correlation. A population of infinitely long messages is considered to be engendered solely by this intersymbol correlation probability: $Q(a_1, .., a_{\nu-1} \mid a_\nu)$ from a given $(\nu - 1)$ -symbol initial sequence. The problem of ergodicity then pertains to existence of unique (i.e., independent of initial sequence), non-vanishing value of $P(a_1, .., a_{\mu-1})$, which should give the probability that a $(\mu - 1)$ -symbol sequence arbitrarily taken from the population is $(a_1, .., a_{\mu-1})$, μ being not necessarily equal to ν. This generalized problem of ergodicity is discussed in our section 2.

It is shown not only that finiteness of correlation range does not warrant ergodicity, as is often erroneously assumed in existing literature, but also that if $\mu < \nu$ the quantity P can have more than one finite value depending on the initial sequence, a situation which does not exist in the ordinary Markoff chains.

Under the conditions that guarantee existence of unique (whether or not non-vanishing) value of P, a convenient quantity, called correlation index W_μ, defined by Eq. (31), is introduced, characterizing both "range" and "strength" of correlation. First, it represents the "range", in the sense that the actual correlation range is the maximum value of μ for which $W_\mu \neq 0$. This criterion is both of theoretical and practical interest. Theoretically, this determines the applicability of the generalized theory of Markoff's chains, and practically, this can be used to measure the existing correlation range in a given population of messages.

Second, this quantity W_μ represents the "strength" of correlation, in the sense that W_μ quantitatively measures the decrease of information due to the existence of μ-symbol correlation as compared with the $(\mu - 1)$-symbol correlation. Finally the so-called redundancy is expressed in the form of a compact series in ascending range-numbers of the correlation indices, Eq. (42).

Ergodicity

We assume the alphabet under consideration to consist of N symbols: $S_1, S_2, \ldots S_N$. We shall constantly use a mathematical symbol:

$$Q(a_1, a_2, \cdots, a_m | a_{m+1}, \cdots, a_{n-1}, a_n), \tag{1}$$

where each one of $a_1, a_2, \ldots, a_n$ can be any one of the N symbols.

Definition I. The quantity denoted by (1) represents the probability that the last (n-m) symbols of a sequence of n symbols are $(a_{m+1}, \ldots, a_n)$ when it is known that the first m symbols of the sequence are $(a_1, \ldots, a_m)$.

By the very nature of probability, we have

$$Q(a_1, \cdots, a_m | a_{m+1}, \cdots, a_n) \geq 0 \quad ; \quad \sum_{a_{m+1}} \cdots \sum_{a_n} Q(a_1, \cdots, a_m | a_{m+1}, \cdots, a_n) = 1 . \tag{2}$$

If there is no correlation between symbols, the probability of any place in a sequence being occupied by symbol S_i is independent of the preceding symbols. As a result, the only quantity which determines a probability of the type (1) is $Q(S_i)$ which represents the probability of symbol S_i appearing at any one place. In this case, we have:

$$Q(a_1, \cdots, a_m | a_{m+1}, \cdots, a_n) = Q(a_{m+1}) Q(a_{m+2}) \cdots Q(a_n) .$$

If the correlation extends, for instance, over three consecutive symbols, and not more than three, then the probability of a place in a sequence being occupied by symbol S_i will depend on the two symbols directly preceding it, but not on the symbols beyond these two. This means that the quantities $Q(S_i, S_j | S_k)$ determine the general probability (1):

$$Q(a_1, \cdots, a_m | a_{m+1}, \cdots, a_n) = Q(a_{m-1}, a_m | a_{m+1}) Q(a_m, a_{m+1} | a_{m+2}) \cdots Q(a_{n-2}, a_{n-1} | a_n) .$$

In general, we have the following theorem:

Theorem I. If the intersymbol correlation does not extend over more than μ consecutive symbols in a sequence, we can factorize (1) as follows:

$$Q(a_1, \cdots, a_m | a_{m+1}, \cdots, a_n) = Q(a_{m-\mu+2}, \cdots, a_m | a_{m+1}) Q(a_{m-\mu+3}, \cdots, a_{m+1} | a_{m+2}) \cdots Q(a_{n-\mu+1}, \cdots, a_{n-1} | a_n) \tag{3}$$

This theorem can be used to define the "range-number" of intersymbol correlation: this number is the minimum allowable μ in the decomposition (3).

Assuming the correlation to be of range ν, we consider all the possible sequences whose first $(\nu - 1)$ symbols are given to be, say, $(a_1, a_2, \ldots, a_{\nu-1})$. Among these sequences starting with $(a_1, a_2, \ldots a_{\nu-1})$, we inquire the probability of those sequences whose first ν symbols are $(a_1, b_1, b_2, \ldots, b_{\nu-1})$. This probability is obviously given by

$$R(a_1, a_2, \cdots, a_{\nu-1} | b_1, b_2, \cdots, b_{\nu-1}) = Q(a_1, a_2, \cdots, a_{\nu-1} | b_{\nu-1}) \quad \text{if} \quad (a_2, \cdots, a_{\nu-1}) = (b_1, \cdots, b_{\nu-2}) ,$$

and otherwise $R(a_1, a_2, \cdots, a_{\nu-1} | b_1, b_2, \cdots, b_{\nu-1}) = 0$.

In other words, the probability in question can be written in a matrix form:

$$(a_1, a_2, \cdots, a_{\nu-1} | R | b_1, b_2, \cdots b_{\nu-1}) = Q(a_1, \cdots, a_{\nu-1} | b_{\nu-1}) \delta(a_2, b_1) \delta(a_3, b_2) \cdots \delta(a_{\nu-1}, b_{\nu-2}) , \tag{4}$$

with

$$\delta(S_i, S_j) = 0 \quad \text{if} \quad i \neq j \quad ; \quad \delta(S_i, S_j) = 1 \quad \text{if} \quad i = j .$$

Using this matrix-expression, the probability, in the above population of sequences, of a particular sequence $(b_1, b_2 \ldots b_{\nu-1})$ appearing in such a position that the place distance between a_1 and b_1 is m symbols can be given by

$$T^{(m)}(a_1, \cdots, a_{\nu-1} | b_1, \cdots, b_{\nu-1}) = (a_1, \cdots, a_{\nu-1} | R^m | b_1, \cdots, b_{\nu-1}) , \tag{5}$$

where R^m simply means the m-th power of R in the sense of matrix-multiplication.

With the help of the quantity (5), we can further calculate the probability of a given sequence of any length $(\mu - 1)$, say $(b_1, \ldots, b_{\mu-1})$, appearing at any position after the initial $(a_1, \ldots, a_{\nu-1})$.

If $\mu > \nu$ this probability will be

$$T^{(m)}(a_1,..,a_{\nu-1}|b_1,..,b_{\mu-1}) = T^{(m)}(a_1,..,a_{\nu-1}|b_1,..,b_{\nu-1})\,Q(b_1,..,b_{\nu-1}|b_\nu)\cdots Q(b_{\mu-2},..,b_{\mu-2}|b_{\mu-1}) \quad (6)$$

where m stands for the symbol distance between a_1 and b_1.

If $\mu < \nu$, we have

$$T^{(m)}(a_1,..,a_{\nu-1}|b_1,..,b_{\mu-1}) = \sum_{b_\mu}\cdots\sum_{b_{\nu-1}} T^{(m)}(a_1,..,a_{\nu-1}|b_1,..,b_{\mu-1},b_\mu,..,b_{\nu-1}), \quad (7)$$

where m bears the same meaning.

Now, the average probability of sequence $(b_1, .., b_{\mu-1})$ with the "place-distance" not larger than m will be

$$U^{(m)}(a_1,..,a_{\nu-1}|b_1,..,b_{\mu-1}) = \frac{1}{m}\sum_{\ell=1}^{m} T^{(\ell)}(a_1,..,a_{\nu-1}|b_1,..,b_{\mu-1}). \quad (8)$$

We now proceed to define what we mean by ergodicity in this paper. We consider all the possible, infinitely long sequences which start with a given initial sequence $(a_1, .., a_{\nu-1})$ and ask the average probability of the sequence $(b_1, .., b_{\mu-1})$ appearing in any position. This probability evidently has the mathematical expression:

$$\lim_{m\to\infty} U^{(m)}(a_1,..,a_{\nu-1}|b_1,..,b_{\mu-1}). \quad (9)$$

The word average here implies a two-fold averaging, viz., first, averaging over all the possible sequences with a fixed position where the sequence $(b_1, .., b_{\mu-1})$ should appear, and second, averaging over all the possible positions of this sequence. The first averaging is mathematically represented by the matrix multiplication in (5), and the second averaging by the summation in (8).

Definition II. If $\lim_{m\to\infty} U^{(m)}(a_1, .., a_{\nu-1}|b_1, .., b_{\mu-1})$ converges to a unique, non-vanishing limit independent of $(a_1, .., a_{\nu-1})$, where $(a_1, .., a_{\nu-1})$ can be taken arbitrarily from a certain family of $(\nu - 1)$-symbol sequences and $(b_1, .., b_{\mu-1})$ can be taken arbitrarily from a certain family of $(\mu - 1)$-symbol sequences, then we speak of ergodicity with regard to these families.

We shall presently see that the quantity (9) with a fixed initial sequence $(a_1, .., a_{\nu-1})$ and a fixed final sequence $(b_1, .., b_{\mu-1})$ indeed converges to a limit, say:

$$U^{(\infty)}(a_1,..,a_{\nu-1}|b_1,..,b_{\mu-1}), \quad (10)$$

but this limit is not necessarily larger than zero, nor is it in general necessarily independent of the initial sequence. In order to understand clearly the situation, let us invoke some well-known mathematical theorems regarding the Markoff chains.[1]

The ordinary Markoff chain formally pertains to a two-symbol correlation probability $(\alpha|R|\beta)$, $(\alpha, \beta = 1, 2, \ldots, M)$: $(\alpha|R|\beta) \geq 1$, $\sum_\beta (\alpha|R|\beta) = 1$. (11)
In accordance with the usual rule of matrix multiplication, we further introduce

$$(\alpha|R^m|\beta) = \sum_\kappa\sum_\lambda\cdots\sum_\mu \underbrace{(\alpha|R|\kappa)(\kappa|R|\lambda)\cdots(\mu|R|\beta)}_{m} \quad (12)$$

Then, we have the following theorems:

Theorem II. The quantity defined by

$$U^{(m)}(\alpha|\beta) = \sum_{\ell=1}^{m}\frac{1}{m}(\alpha|R^\ell|\beta) \quad (13)$$

for any given pair (α, β) converges to a limit as $m\to\infty$:

$$U^{(\infty)}(\alpha|\beta) = \lim_{m\to\infty} U^{(m)}(\alpha|\beta). \quad (14)$$

Theorem III. The entire set G of symbols $(\alpha = 1,2, .., M)$ can be divided into a "vanishing" subset V and a certain number of "closed" subsets $C_i (i = 1,2, ..)$ in such a way that

$U^{(\infty)}(\alpha|\beta) = 0$ for α belonging to G, and for β belonging to V,
$U^{(\infty)}(\alpha|\beta) > 0$ for α and β belonging to the same C_i,
$U^{(\infty)}(\alpha|\beta) = 0$ for α and β belonging to different C's.

Theorem IV. $U^{(\infty)}(\alpha|\beta)$ is independent of α, if α and β belong to the same C.

Coming back to our original topic, if the correlation-range is two, and if $\mu=\nu$, these theorems can be directly applied to our problem involved in Definition II. If the correlation-range is > 2, we only need to consider a sequence of $(\nu-1)$ symbols collectively as a symbol α. The R's defined in (4) indeed satisfy (11). The cases: $\mu\neq\nu$ can be handled with the help of (6) and (7).

From Theorem II follows quite generally:

Theorem V. The limit (10) exists.

We shall now discuss first the case $\mu=\nu$ in the light of Theorems II, III and IV. According to Theorem III, the entire set of $(\nu-1)$-symbol sequences is subdivided into a vanishing subset V and a certain number of closed subsets C_i. If the final sequence of (10) belongs to V, then $U^{(\infty)}$ is zero independently of the initial sequence. For a given final sequence belonging to one of the closed subsets, $U^{(\infty)}$ will be zero if the initial sequence belongs to another closed subset, and will have a constant non-vanishing value insofar as the initial sequence belongs to the same closed subset as the final sequence. Thus:

Theorem VI. When $\mu=\nu$, ergodicity in the sense of Def. II holds if and only if the initial family and the final family are the same closed subset.

In the cases where $\mu>\nu$, we construct an "extended" closed subset D_i of $(\mu-1)$ symbols by taking those $(\mu-1)$-symbol sequences $(b_1, .., b_{\mu-1})$ whose first $(\nu-1)$ symbols coincide with one of the members of the $(\nu-1)$-symbol closed subset C_i and which satisfy the condition:

$$Q(b_1, .., b_{\nu-1}|b_\nu)\,Q(b_2, .., b_\nu|b_{\nu+1})\cdots Q(b_{\mu-\nu}, .., b_{\mu-2}|b_{\mu-1})\neq 0. \qquad (15)$$

The extended vanishing subset will be composed of all those $(\mu-1)$-symbol sequences whose first $(\nu-1)$ symbols coincide with one of the members of the $(\nu-1)$-symbol vanishing subset, or whose first $(\nu-1)$ symbols coincide with one of the members of some closed subset but whose last $(\mu-\nu)$ symbols violate the condition (15). The entire set of possible $(\mu-1)$-symbol sequences are thus covered by the D's and V, and there is no possible overlapping. If the $(\mu-1)$-symbol final sequence of (10) is a member of this extended vanishing subset, $U^{(\infty)}$ will certainly vanish whatever the initial sequence may be. If the final sequence belongs to an extended closed subset D_i, then $U^{(\infty)}$ will vanish for an initial sequence belonging to a C_j different from the one, C_i, which corresponds to D_i, and will have a constant non-vanishing value for any initial sequence belonging to C_i.

Theorem VII. When $\nu<\mu$, ergodicity holds if and only if the initial family is one of the closed subset C_i and the final family is the extended closed subset D_i corresponding to C_i.

In the cases where $\mu<\nu$, we encounter a rather peculiar situation. From a closed subset C_i we construct a retrenched subset E_i of $(\mu-1)$-symbol sequences. E_i is the set of those $(\mu-1)$ symbols sequences which coincide with the first $(\mu-1)$ symbols of at least one of the members of C_i. The retrenched vanishing subset is defined as the totality of all those $(\mu-1)$-symbol sequences which do not belong to any one of the retrenched closed subsets. In case of the extended closed subsets, a given sequence of $(\mu-1)$ symbols could not belong to more than one D_i, since the division made in Theorem III does not allow for any overlapping. However, in the present case of retrenched subsets, a given $(\mu-1)$-symbol sequence may well belong to more than one E. If the $(\mu-1)$-symbol final sequence of (10) belongs to the retrenched vanishing subset, $U^{(\infty)}$ will always vanish. If the $(\mu-1)$-symbol final sequence belongs to $E_i, E_j, .., E_k$, then $U^{(\infty)}$ will be zero for an initial sequence belonging to a C different from any one of the corresponding subsets: $C_i, C_j, .., C_k$. For the same final sequence $U^{(\infty)}$ may thus have different non-vanishing values according as to which one of $C_i, C_j, .., C_k$ the initial sequence belongs.

Theorem VIII. When $\mu<\nu$, ergodicity holds for the initial family identical with one of the closed subset C_i and the final family identical with the corresponding retrenched subset E_i.

In the foregoing considerations, we have systematically omitted the initial sequences belonging to the vanishing subset V. The reason for this is that the $U^{(\infty)}$ depends in this case on the detailed structure of the intersymbol correlation, and that we cannot draw a conclusion of general validity. (Of course, if the final sequence also belongs to V, then $U^{(\infty)}$ vanishes).

Regarding the closed subsets of $(\nu-1)$ symbols, we should like to mention the following interesting property. We have obviously

$$U^{(\infty)}(a_1,\cdots,a_{\nu-1}\mid b_2,\cdots,b_\nu)=\sum_{b_1}U^{(\infty)}(a_1,\cdots,a_{\nu-1}\mid b_1,\cdots,b_{\nu-1})\,Q(b_1,\cdots,b_{\nu-1}\mid b_\nu),$$

whence we infer:

Theorem IX. $(b_2, b_3, \cdots b_\nu)$ is a member of C_i, if there is any symbol b_1 such that $(b_1, b_2, \cdots, b_{\nu-1})$ is a member of C_i and $Q(b_1, b_2, \cdots, b_{\nu-1} \mid b_\nu) \neq 0$.

For a given $(b_1, b_2, \cdots, b_{\nu-1})$ there must be at least one b_ν such that $Q(b_1, b_2, \cdots, b_{\nu-1} \mid b_\nu) \neq 0$, on account of (2). Hence:

Theorem X. If $(b_1, b_2, \cdots, b_{\nu-1})$ is a member of C_i, then there is always a member of C_i whose first $(\nu - 2)$ symbols are $(b_2, \cdots, b_{\nu-1})$.

Before closing this section, a simple illustration may be given. Suppose the alphabet to be composed of three symbols: S_1, S_2 and S_3, and to have an intersymbol correlation of range 3:

$$Q(S_1, S_1 \mid S_1) = 1, \quad Q(S_1, S_2 \mid S_1) = 1,$$
$$Q(S_1, S_3 \mid S_1) = 1, \quad Q(S_2, S_1 \mid S_2) = 1,$$
$$Q(S_2, S_2 \mid S_1) = 1, \quad Q(S_2, S_3 \mid S_1) = 1,$$
$$Q(S_3, S_1 \mid S_1) = 1, \quad Q(S_3, S_2 \mid S_1) = 1,$$
$$Q(S_3, S_3 \mid S_1) = 1.$$

Then the $(\nu - 1)$ symbol subsets are:

C_1 : (S_1, S_1)
C_2 : (S_1, S_2), (S_2, S_1)
C_3 : (S_2, S_2)
V : (S_1, S_3), (S_3, S_1), (S_2, S_3), (S_3, S_2), (S_3, S_3)

The extended 3-symbol subsets are:

D_1 : (S_1, S_1, S_1)
D_2 : (S_1, S_2, S_1), (S_2, S_1, S_2)
D_3 : (S_2, S_2, S_2)
V' : all other 3-symbol sequences

The retrenched 1-symbol subsets are:

E_1 : S_1
E_2 : S_1, S_2
E_3 : S_2
V' : S_3

We can see the overlapping we have discussed; as a result, $U^{(\infty)}$ with the final sequence (symbol) S_1, for instance, becomes three-valued:

$U^{(\infty)}(S_1, S_1 \mid S_1) = 1$
$U^{(\infty)}(S_1, S_2 \mid S_1) = \frac{1}{2}$
$U^{(\infty)}(S_2, S_1 \mid S_1) = \frac{1}{2}$
$U^{(\infty)}(S_2, S_2 \mid S_1) = 0$
All other $U^{(\infty)}(\mid S_1) = 1$

Redundancy

In this section, we shall constantly use a quantity denoted by:

$$P(a_1, a_2, .., a_n) \geq 0 . \tag{17}$$

Definition III. The quantity (17) represents the probability, in infinitely long messages, of an arbitrarily taken sequence of symbol-length n being a particular sequence $(a_1, a_2, .., a_n)$.

From this definition follows the normalization condition:

$$\sum_{a_1} \cdots \sum_{a_n} P(a_1, a_2, .., a_n) = 1 . \tag{18}$$

According to the point of view of the last section, the existence of a unique value of such a probability is not unconditionally guaranteed. Only if the initial sequence $(b_1, .., b_{\nu-1})$ is limited to within a closed subset, say, C_i, then

$$U^{(\infty)}(b_1, .., b_{\nu-1} \mid a_1, .., a_n)$$

becomes independent of $(b_1, .., b_{\nu-1})$, i.e., a function only of $(a_1, .., a_n)$. If this is the case, we can write

$$U^{(\infty)}(b_1, .., b_{\nu-1} \mid a_1, .., a_n) = P(a_1, .., a_n) . \tag{19}$$

According to the theorems of the last section, if $(a_1, .., a_n)$ belongs to C_i, or its extended subset D_i, or its retrenched subset E_i, P will be finite, and otherwise zero. We have therefore to restrict the "infinitely long messages" of Definition III to only those which start with initial sequences belonging to one closed subset. The condition regarding P does not require that all the P's should be non-vanishing, thence the restriction on the final sequences, in the sense of Definition II, is not necessary. On account of ergodicity, two sequences starting from two different initial sequences of the same closed subset becomes, in the long run, statistically identical. It is true that we can evade the restriction on the initial sequences by giving a certain "weight" to each of the closed subsets, which would lead to a unique value of each P. However, from the point of view that the messages are engendered solely by the correlation probability, this alternative is not acceptable, since it involves an arbitrary "weight" of each closed subset. Our discussion of this section will be based on the assumption that the initial sequences are limited to a single subset. The generalization of the results to the case of "weighted" subsets is very simple.

It should be noted that, as a result of the limitation of the initial sequences to a single subset, it may well happen that some of the generally possible sequences $(a_1, .., a_{\nu-1})$ in the correlation probability $Q(a_1, .., a_{\nu-1} \mid a_\nu)$ actually never happen in the possible messages. Thus the actual range of correlation may become smaller than the range defined with regard to the entire possibilities of the a's. For instance, in the illustration of the last section, if we limit ourselves to the initial subset C_2, all 3-symbol Q's except $Q(S_1, S_2 \mid S_1) = 1$ and $Q(S_2, S_1 \mid S_2) = 1$ will become meaningless. These two 3-symbol correlation probabilities reduce to the following two 2-symbol correlation probabilities: $Q(S_1 \mid S_2) = 1$, and $Q(S_2 \mid S_1) = 1$. The range is thus reduced from three to two.

In the empirical point of view, if a population of very long sample messages is given, we can always evaluate (17) by just counting the frequency of each segment $(a_1, .., a_n)$. However, if we divide this entire population into, say, two groups, the values of (17) may be different in the two groups. This discrepancy may be caused by a difference in correlation probabilities and/or by a difference in the initial sequences. We thus see that the problem of ergodicity is not irrelevant to the empirical point of view. In this section, however, we assume that we have a single population from which the quantities of the type (17) are uniquely determined.

The quantity (17) has, besides (18), the property:

$$\sum_{a} P(a_1, .., a_k, b_1, \cdots, b_m, a_{k+m+1}, .., a_n) = P(b_1, \cdots, b_m) . \tag{20}$$

This is obvious from the statistical point of view, but can also be verified from the standpoint of (19).

According to (6), we have for $n \geq \nu$

$$P(a_1, .., a_n) = P(a_1, .., a_{\nu-1}) Q(a_1, .., a_{\nu-1} \mid a_\nu) \cdots Q(a_{n-\nu+1}, .., a_{n-1} \mid a_n) , \tag{21}$$

or more generally,

$$P(a_1,..,a_n) = P(a_1,..,a_{\mu-1})\,Q(a_1,..,a_{\mu-1}|a_\mu)\cdots Q(a_{n-\mu+1},..,a_{n-1}|a_n), \tag{22}$$

provided $n \geq \mu \geq \nu$. Equivalence of (21) and (22) can readily be seen with the help of (3) and (6). In particular, for $n=\mu\geq\nu$, we get from (22)

$$Q(a_1,..,a_{\mu-1}|a_\mu) = \frac{P(a_1,..,a_\mu)}{P(a_1,..,a_{\mu-1})} \quad . \tag{23}$$

This is just what should be according to Definitions I and III. (23) may be considered as the definition of $Q(a_1, .., a_{\mu-1}|a_\mu)$ even for $\mu<\nu$. However, with such Q's with $\mu<\nu$, (22) will not be true, since the Q's with $\mu<\nu$ cannot describe fully the existing correlation.

Substituting (23) into (22), we get

$$P(a_1,..,a_n) = \frac{P(a_1,..,a_\mu)P(a_2,..,P_{\mu+1})\cdots P(a_{n-\mu+1},..,a_n)}{P(a_2,..,a_\mu)\cdots P(a_{n-\mu+1},..,a_{n-1})} \quad , \tag{24}$$

provided $n>\mu\geq\nu$. The actual range ν is thus the minimum value of μ for which the decomposition (24) is allowed.

For an allowed value of μ, if a further decomposition of range $\mu-1$ is still allowed, i.e., if $\mu-1\geq\nu$, then we get from (24)

$$P(a_1,..,a_\mu) = \frac{P(a_1,..,a_{\mu-1})\,P(a_2,..,a_\mu)}{P(a_2,..,a_{\mu-1})} \tag{25}$$

for all $(a_1, .., a_\mu)$. But if $\mu-1<\nu$, the left side of (25) will not be equal to its right side for at least one sequence $(a_1, .., a_\mu)$. Thus we are led to use (25) as a criterion to determine whether $\mu>\nu$ or not: If (25) holds for all $(a_1, .., a_\mu)$, then $\mu>\nu$; if not, $\mu\leq\nu$. Indeed, if (25) is possible, we have in virtue of (23),

$$Q(a_1,..,a_{\mu-1}|a_\mu) = \frac{P(a_1,..,a_\mu)}{P(a_1,..,a_{\mu-1})} = \frac{P(a_2,..,a_\mu)}{P(a_2,..,a_{\mu-1})} = Q(a_2,..,a_{\mu-1}|a_\mu) \quad , \tag{26}$$

i.e., Q of range μ is reducible to a Q of range $(\mu-1)$. In the light of Theorem I, this means that the actual range is $(\mu-1)$ or less. If (25) breaks down for at least one sequence $(a_1, .., a_\mu)$, then (26) does not hold in general, meaning that the actual range is larger than $(\mu-1)$.

Theorem XI. If and only if (25) holds for all $(a_1, .., a_\mu)$, the actual correlation range ν is $(\mu-1)$ or less.

This criterion is interesting particularly in the empirical point of view, for here the P's, instead of the Q's, are the quantities which are primarily given. The criterion of Theorem XI can be brought to a more concise form by the help of the well-known theorem attributed to W. Gibbs:

Theorem XII. If

$$f_i \geq 0\,,\ g_i \geq 0\,, \quad \text{and} \quad \sum_i f_i = \sum_i g_i\,, \ (i=1,2,\cdots,r), \tag{27}$$

then

$$W \equiv \sum_i f_i \log f_i - \sum f_i \log g_i \geq 0\,, \tag{28}$$

where the equality holds only when $f_i = g_i$ for all i.

Now, let us call the left-hand side and the right-hand side of (25), respectively

$$f(a_1,..,a_\mu) = P(a_1,..,a_\mu) \tag{29}$$

$$g(a_1,..,a_\mu) = \frac{P(a_1,..,a_{\mu-1})\,P(a_2,..,a_\mu)}{P(a_2,..,a_{\mu-1})}\,, \tag{30}$$

and consider the index i of Theorem XII as a collective index for various possible sequences of symbol-length μ. On account of (18) and (20), the conditions (27) are satisfied, and we obtain

$$W_\mu \equiv \sum P(a_1,..,a_\mu)\log P(a_1,..,a_\mu) - 2\sum P(a_1,..,a_{\mu-1})\log(a_1,..,a_{\mu-1}) + \sum P(a_1,..,a_{\mu-2})\log(a_1,..,a_{\mu-2}) \geq 0\,. \tag{31}$$

Only when (25) holds for all $(a_1, .., a_\mu)$, then $W_\mu = 0$. In other words, for a given value of ν, $W_\mu = 0$ for $\mu>\nu$. This leads to a convenient way to determine the actual range:

Theorem XIII. The actual range ν is the maximum value of μ for which $W_\mu \neq 0$.

The W's defined by (31) will be called "correlation indicies".

For $\mu = 2$, the definition of W_μ in (31) should be understood as meaning

$$W_2 = \sum P(a_1,a_2) \log P(a_1,a_2) - 2 \sum P(a_1) \log P(a_1) , \tag{32}$$

for we have here $g(a_1, a_2) = P(a_1)P(a_2)$.

We shall now proceed to find out the average amount of information carried by a message-segment of length n in a language in which the P's exist. A specific message-segment $(a_1, .., a_n)$ has probability $P(a_1, .., a_n)$. Thus the information per symbol carried by this message-segment is

$$-\frac{1}{n} \log P(a_1, .., a_n).$$

The probability of occurrence of such a message being $P(a_1, .., a_n)$, the average information per symbol for various possible message-segments of length n is given by

$$I_n = -\frac{1}{n} \sum P(a_1, .., a_n) \log P(a_1, .., a_n) . \tag{33}$$

Now, if the existing correlation is of range ν, the P can be decomposed as in (24) with $\mu = \nu$. A straightforward calculation with the help of (18) and (20) gives

$$I_n = I_{n,\nu} \equiv -\frac{1}{n}(n-\nu+1) \sum P(a_1, .., a_\nu) \log P(a_1, .., a_\nu) + \frac{1}{n}(n-\nu) \sum P(a_1, .., a_{\nu-1}) \log P(a_1, .., a_{\nu-1}) . \tag{34}$$

For an obvious reason this ν can be the actual minimum range or any ν that is larger than this. Supposing ν in (34) to be the actual minimum range, let us find the error which would be committed by the calculation based on the assumption that the actual range were $\nu - 1$. This is easily found to be

$$I_{n,\nu} - I_{n,\nu-1} = -\frac{(n-\nu+1)}{n} W_\nu . \tag{35}$$

Repeating this process, we obtain

$$I_n - I^\circ = I_{n,\nu} - I^\circ = -\sum_{\mu=2}^{\nu} \frac{n-\mu+1}{n} W_\mu , \tag{36}$$

where

$$I^\circ \equiv I_{n,1} = -\sum P(a_1) \log P(a_1) . \tag{37}$$

Since W_μ vanishes anyway for $\mu > \nu$, we can state:

Theorem XIV. The average information per symbol carried by a message-segment of length n is

$$I_n = I^\circ - \sum_{\mu=2}^{\infty} \frac{n-\mu+1}{n} W_\mu \tag{38}$$

insofar as n is larger than the actual correlation range.

Since the W's are zero or positive, the intersymbol correlation tends to decrease the amount of information. Thus, W_μ can be considered to represent the "strength" of correlation — strength in the sense of reducing the amount of information. By definition, I_n cannot be negative, thence there is an upper limit to the total "strength" of the correlation:

$$\sum_{\mu=2}^{\infty} \frac{n-\mu+1}{n} W_\mu \le \sum_{\mu=2}^{\infty} W_\mu \le I^\circ . \tag{39}$$

For $n \gg \nu$, we obtain from (38)

$$I_n \approx I_\infty = I^\circ - \sum_{\mu=2}^{\infty} W_\mu , \quad (n \gg \nu) , \tag{40}$$

showing that if we take a sufficiently long segment as a unit, the information per symbol becomes independent of the length of the segment. This indirectly justifies the usual procedure according to which an infinitely long message is cut into segments of sufficient length and the segments are treated as if they did not have any correlation among them.

The quantity called "redundancy" is defined by[2]

$$R = (I^\circ - I_\infty)/I^\circ . \tag{41}$$

Theorem XV. The redundancy of a language which is characterized by the correlation indices W_μ is given by

$$R = (1/I^\circ) \sum_\mu W_\mu , \qquad 0 \le R \le 1 . \tag{42}$$

In the illustration of the last section, if we limit the initial sequences to C_2, we get

$$W_2 = \log 2 , \quad W_3 = W_4 = \cdots = 0 , \quad I^\circ = \log 2 , \quad I_\infty = 0 , \quad R = 100\%$$

This last result is not surprising, because the possible infinite sequences are limited to: $..S_1S_2S_1S_2..$, which certainly cannot convey any information.

1. See for instance W. Feller, Introduction to Probability Theory and its Applications. (Wiley, N.Y. 50) p 307 ff.
2. Stanford Goldman, Information Theory (Prentice-Hall, N.Y., 53) p 45.

MULTIVARIATE INFORMATION TRANSMISSION*

William J. McGill**

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

RESEARCH LABORATORY OF ELECTRONICS

and

LINCOLN LABORATORY

ABSTRACT

A multivariate analysis based on transmitted information is presented. It is shown that sample transmitted information provides a simple method for measuring and testing association in multidimensional contingency tables. Relations with analysis of variance are pointed out, and statistical tests are described.

*The research in this document was supported jointly by the Army, Navy, and Air Force under contract with the Massachusetts Institute of Technology.

**Staff Member, Lincoln Laboratory, Massachusetts Institute of Technology.

MULTIVARIATE INFORMATION TRANSMISSION*

I. INTRODUCTION

Several recent articles in the psychological journals have shown how ideas derived from communication theory are being applied in psychology. It is not widely understood, however, that the tools made available by communication theory are useful for analyzing data, whether or not we believe that the human organism is best described as a communications system.

This memorandum will present an extension of Shannon's[10] measure of transmitted information. It will be shown that transmitted information leads to a simple multivariate analysis of contingency data and to appropriate statistical tests.

II. BASIC DEFINITIONS

Let us consider a communication channel and its input and output. Transmitted information measures the amount of association between the input and the output of the channel. If input and output are perfectly correlated, all the input information is transmitted. On the other hand, if input and output are independent, no information is transmitted. Naturally, most cases of information transmission are found between these extremes. There is some uncertainty at the receiver about what was sent. Some information is transmitted and some does not get through.

We are interested not in what the transmitted information is, but in the amount of information transmitted. Suppose that we have a discrete input variable, x, and a discrete output variable, y. Since x is discrete, it takes on values or signals k = 1,2,3, ... X with probabilities indicated by p(k). Similarly, y assumes values m = 1,2,3, ... Y with probabilities p(m). If it happens that k is sent and m is received, we can speak of the joint input-output event (k,m). This joint event has probability p(k,m). The rules governing the selection of signals at either end of the channel must be constructed so that

$$\sum_{k=1}^{k=X} p(k) = \sum_{m=1}^{m=Y} p(m) = \sum_{k,m} p(k,m) = 1 \quad .$$

Under these conditions, and if successive signals are independent, the amount of information transmitted in "bits" per signal is defined as

$$T(x;y) = H(x) + H(y) - H(x,y) \quad , \tag{1}$$

where

$$H(x) = -\sum_{k} p(k) \log_2 p(k) \quad ,$$

$$H(y) = -\sum_{m} p(m) \log_2 p(m) \quad ,$$

$$H(x,y) = -\sum_{k,m} p(k,m) \log_2 p(k,m) \quad .$$

One "bit" is equal to $-\log_2 (1/2)$ and represents the information conveyed by a choice between two equally probable alternatives. Our development will use the bit as a unit, since this is the

*Several of the indices and tests discussed in this paper have been developed independently by J. E. Keith Smith at the University of Michigan, and by W. R. Garner at Johns Hopkins University.

convention in information theory, but any convenient unit may be substituted by changing the base of the logarithm.

If there is a relation between x and y, $H(x) + H(y) > H(x,y)$ and the size of the inequality is just $T(x;y)$. On the other hand, if x and y are independent, $H(x,y) = H(x) + H(y)$ and $T(x;y)$ is zero. It can be shown that $T(x;y)$ is never negative.

The presentation to this point has been an outline of the properties of the measure of transmitted information as set forth by Shannon.[10] These properties may be summarized by stating that the amount of information transmitted is a bivariate, positive quantity that measures the association between input and output of a channel. There are, however, very few restrictions on how a channel may be defined. The input-output relations that occur in many psychological contexts are certainly possible channels. Consequently, we can measure transmitted information in these contexts and anticipate that the results will be interesting.

III. SAMPLE INFORMATION

Our development will be based on sample measures of information, i.e., on measures of information constructed from relative frequencies.

Suppose that we make n observations of events (k,m). We identify n_{km} as the number of times that k was sent and m was received. This means that

$$n_k = \sum_m n_{km} \quad ,$$

$$n_m = \sum_k n_{km} \quad ,$$

$$n = \sum_{k,m} n_{km} \quad ,$$

where n_k is the number of times that k was sent, n_m is the number of times that m was received, and n is the total number of observations. A particular experiment can then be represented by a contingency table with XY cells and entries n_{km}.

We may estimate the probabilities p(k), p(m) and p(k,m) with n_k/n, n_m/n and n_{km}/n respectively. Sample transmitted information $T'(x;y)$ is defined as*

$$T'(x;y) = H'(x) + H'(y) - H'(x,y) \quad , \tag{2}$$

where $H'(x)$, $H'(y)$ and $H'(x,y)$ are constructed from relative frequencies instead of from probabilities. As before, $T'(x;y)$ is the amount of transmitted information (in the sample) measured in "bits" per signal.

Since it is difficult to manipulate logs of relative frequencies, we will introduce an easier notation:

$$s_{km} = \frac{1}{n} \sum_{k,m} n_{km} \log_2 n_{km} \quad ,$$

*Throughout this memorandum, a prime is used over a quantity to indicate the maximum likelihood estimator of the same quantity without the prime. For example, $T'(u;y)$ is an estimator for $T(u;y)$.

$$s_k = \frac{1}{n} \sum_k n_k \log_2 n_k \quad ,$$

$$s_m = \frac{1}{n} \sum_m n_m \log_2 n_m \quad ,$$

$$s = \log_2 n \quad .$$

Expressions involving sample measures of information are easier to handle in this notation. For example, $T'(x;y)$ becomes

$$T'(x;y) = s - s_k - s_m + s_{km} \quad . \tag{3}$$

Equations (2) and (3) are equivalent expressions for $T'(x;y)$. When we write equations like (3), we shall say that these equations are written in s-notation. Thus Eq. (3) is Eq. (2) in s-notation.

IV. THREE-DIMENSIONAL TRANSMITTED INFORMATION

Now let us extend the definition of transmitted information to include two sources, u and v, that transmit to y. To accomplish this, we replace x in Eq. (2) with u, v and we find that

$$T'(u,v;y) = H'(u,v) + H'(y) - H'(u,v,y) \quad , \tag{4}$$

where x has been subdivided into two classes, u and v. The possible values of u are $i = 1, 2, 3, \ldots U$, while v assumes values $j = 1, 2, 3, \ldots V$. The subdivision is arranged so that the range of values of u and v jointly constitute the possible values of x. This means that the input event k can be replaced by the joint input event (i, j). Consequently, we have

$$n_k = n_{ij} \quad ,$$

and the direct substitution of u, v for x in Eq. (2) is legitimate.

Our new term, $T'(u,v;y)$, measures the amount of information transmitted when u and v transmit to y. It is evident, however, that the direction of transmission is irrelevant, for examination of Eq. (4) reveals that

$$T'(u,v;y) = T'(y;u,v) \quad .$$

This means that nothing is gained formally by distinguishing transmitters from receivers. The amount of information transmitted is a measure of association between variables. It does not respect the direction in which the information is traveling. On the other hand, we cannot permute symbols at will, for

$$T'(u,y;v) = H'(u,y) + H'(v) - H'(u,v,y) \quad ,$$

and this is not necessarily equal to $T'(u,v;y)$.

Our aim now is to measure $T'(u,v;y)$, and then to express $T'(u,v;y)$ as a function of the bivariate transmissions between u and y, and v and y. Computation of $T'(u,v;y)$ is not difficult. Our observations of the joint event (i, j, m) organize themselves into a three-dimensional

contingency table with UVY cells and entries n_{ijm}. We can compute the quantities in Eq. (4) from this table, or we can write

$$T'(u,v;y) = s - s_m - s_{ij} + s_{ijm} \quad , \tag{5}$$

where

$$s_{ijm} = \frac{1}{n} \sum_{i,j,m} n_{ijm} \log_2 n_{ijm} \quad ,$$

and other s-terms are defined by analogy with the s-terms in Eq. (3).

Now suppose that we want to study transmission between u and y. We may eliminate v in two ways. First, let us reduce the three-dimensional contingency table to two dimensions by summing over v. The entries in the reduced table are

$$n_{im} = \sum_j n_{ijm} \quad .$$

We have, for the transmitted information between u and y,

$$T'(u;y) = s - s_i - s_m + s_{im} \quad . \tag{6}$$

The second way to eliminate v is to compute the transmission between u and y separately for each value of v, and then average these together. This transmitted information will be called $T'_v(u;y)$, where

$$T'_v(u;y) = \sum_j \frac{n}{n} j \, [T'_j(u;y)] \quad , \tag{7}$$

and $T'_j(u;y)$ is information transmitted between u and y for a single value of v, namely, j. It is readily shown that

$$T'_v(u;y) = s_j - s_{ij} - s_{jm} + s_{ijm} \quad . \tag{8}$$

We see that $T'_v(u;y)$ is written in the same way as $T'(u;y)$, except that the subscript j is added to each of the s-terms.

There are three different pairs of variables in a three-dimensional contingency table. For example, the two equations for transmission between v and y are written

$$T'(v;y) = s - s_j - s_m + s_{jm} \quad , \tag{9}$$

$$T'_u(v;y) = s_i - s_{ij} - s_{im} + s_{ijm} \quad . \tag{10}$$

Finally, we may study transmission between u and v, i.e.,

$$T'(u;v) = s - s_i - s_j + s_{ij} \quad , \tag{11}$$

$$T'_y(u;v) = s_m - s_{im} - s_{jm} + s_{ijm} \quad . \tag{12}$$

With these results in mind, let us reconsider the information transmitted between u and y. If v has an effect on transmission between u and y, then $T'_v(u;y) \neq T'(u;y)$. One way to measure the size of the effect is by

$$A'(uvy) = T'_v(u;y) - T'(u;y) \quad ,$$

$$A'(uvy) = -s + s_i + s_j + s_m - s_{ij} - s_{im} - s_{jm} + s_{ijm} \quad . \tag{13}$$

A few more substitutions will show that

$$\begin{aligned} A'(uvy) &= T'_v(u;y) - T'(u;y) \quad , \\ &= T'_u(v;y) - T'(v;y) \quad , \\ &= T'_y(u;v) - T'(u;v) \quad . \end{aligned} \tag{14}$$

In view of this symmetry, we may call A'(uvy) the u . v . y interaction information. We see that A'(uvy) is the gain (or loss) in sample information transmitted between any two of the variables, due to additional knowledge of the third variable.

Now we can express the three-dimensional information transmitted from u, v to y, i.e., T'(u, v;y), as a function of its bivariate components, for

$$T'(u,v;y) = T'(u;y) + T'(v;y) + A'(uvy) \quad , \tag{15}$$

$$T'(u,v;y) = T'_v(u;y) + T'_u(v;y) - A'(uvy) \quad . \tag{16}$$

Equations (15) and (16) taken together mean that T'(u, v;y) can be represented by a diagram with overlapping circles as shown in Fig. 1. The diagram assumes what we shall call "positive" interaction between u, v and y. Interaction is positive when the effect of holding one of the interacting variables constant is to increase the amount of association between the other two. This means that $T'_v(u;y) > T'(u;y)$, and $T'_u(v;y) > T'(v;y)$. (Because of Eq. (14), if one of these inequalities holds, both must hold.) Later on, however, we shall show that interaction may be negative. When this happens, relations between the interacting variables are reversed, and the diagram in Fig. 1 is no longer strictly correct.

Fig. 1. Schematic diagram of the components of three-dimensional transmitted information. The diagram shows that three-dimensional transmission can be analyzed into a pair of bivariate transmissions plus an interaction term. The meanings of the symbols are explained in the text.

V. COMPONENTS OF RESPONSE INFORMATION

The multivariate model of information transmission is useful to us because the situations treated by communication theory are not the same as those with which we deal in psychological applications. The engineer is usually able to restrict himself to transmission from a single information source. He knows the statistical properties of the source, and when he speaks of noise he means random noise. This kind of precision is seldom available to us. In our experiments we generally do not know in advance how many sources are transmitting information. We must therefore be careful not to confuse statistical noise with the experimenter's ignorance.

The bivariate model of transmitted information provided by communication theory tells us to attribute to random noise whatever uncertainty there is in specifying the response when

the stimulus is known.[1] Consequently, if several sources transmit information to responses, the bivariate model will certainly fail to discriminate effects due to uncontrolled sources from those due to random variability. On the other hand, the multivariate model can measure the effects due to the various transmitting sources. For example, in three-dimensional transmission we find that

$$H'(y) = H'_{uv}(y) + T'(u;y) + T'(v;y) + A'(uvy) \quad , \tag{17}$$

where $H'(y) = s - s_m$ and $H'_{uv}(y) = s_{ij} - s_{ijm}$.

We see that H'(y), the response information, has been analyzed into an error term plus a set of correlation terms due to the input variables. The error term $H'_{uv}(y)$ is the residual or unexplained variability in the output y after the information due to the inputs u and v has been removed. In bivariate information transmission, the response information is analyzed less precisely. For example, we may have

$$H'(y) = H'_u(y) + T'(u;y) \tag{18}$$

In this case, the error term is $H'_u(y)$ because only one input, u, is recorded. Shannon[10] showed that

$$H'_u(y) \geq H'_{uv}(y) \quad .$$

In other words, the error term, when only u is controlled, cannot be increased if we also control v. In fact,

$$H'_u(y) = H'_{uv}(y) + T'_u(v;y) \quad . \tag{19}$$

Equation (19) is proved by expanding both sides in s-notation. Thus, if u and v are stimulus variables that transmit information via responses y, we have an error term $H'_u(y)$, provided we keep track of only one of the inputs, namely, u. However, this error term contains a still smaller error term, as well as the information transmitted from v. Controlling v is thus seen to be equivalent to extracting the association between v and y from the noise. Multivariate transmitted information is essentially information analyzed from the noise part of bivariate transmission.

VI. AN EXAMPLE

The kind of analysis that multivariate information transmission yields can be illustrated by a set of data obtained from one subject in an experiment on frequency judgment.

Four equally loud tones – 890, 925, 970 and 1005 cycles per second – were presented to the subject one at a time in random order. Each tone was 1/2 second long and separated by about 3 seconds from the next tone. During preliminary training, the subject learned to identify the tones by pairing them with four response keys. In experimental sessions, a loud masking noise was turned on, and a random sequence of 250 tones was presented against the noise background. A flashing light told the subject when the stimulus occurred, and he was instructed to guess, if in doubt, about which one of the four tones it was.

One object of the experiment was to find weights for both the frequency stimulus and the immediately preceding response in determining which key the subject would press. Tests were run at several signal-to-noise ratios. The data presented here were obtained when the signal-to-noise ratio was close to the masked threshold.

In order to calculate weights, we can consider the experiment as an example of three-dimensional transmission. Our analysis is based on the responses to the 125 even-numbered stimuli. The odd-numbered responses are considered as the context in which the subject judged the even-numbered stimuli. The odd-numbered stimuli are ignored in this analysis.

The stimuli will be designated as the variable u. Last previous responses are called "presponses" and they will be indicated by the variable v. These are the inputs. Current responses are represented by y. This is the output variable. Thus we can identify the joint event (i,j,m) as the occurrence of response m to stimulus i, following presponse j. Failure to respond is considered as a possible response. Consequently, there are four stimulus categories and five response categories.

The subject's responses to the 125 test stimuli were sorted into a $4 \cdot 5 \cdot 5$ contingency table. Two of the reduced tables that were obtained from this master table are reproduced here in order to illustrate our computations. For example, the stimulus-response plot in Table I has entries n_{im}. The calculation for s_{im} goes as follows:

$$s_{im} = \frac{1}{125} [1 \log_2 1 + 5 \log_2 5 + 12 \log_2 12 + \ldots\ldots + 7 \log_2 7 + 10 \log_2 10] \quad ,$$

$$s_{im} = 374.05750/125 \quad ,$$

$$s_{im} = 2.99246 \quad .$$

In the same way, s_{jm} is computed from the figures for n_{jm} in the presponse-response table (Table II).

TABLE I

STIMULUS-RESPONSE FREQUENCY TABLE

Response \ Stimulus	1	2	3	4	
0	1	3	2	1	7
1	5	2	2	1	10
2	12	10	13	12	47
3	8	10	12	7	37
4	5	5	4	10	24
	31	30	33	31	125

TABLE II

PRESPONSE-RESPONSE FREQUENCY TABLE

Response \ Presponse	0	1	2	3	4	
0	1	2	3	0	1	7
1	1	1	4	3	1	10
2	2	13	8	20	4	47
3	3	7	12	6	9	37
4	3	3	0	15	3	24
	10	26	27	44	18	125

$$s_{jm} = \frac{1}{125} [1 \log_2 1 + 1 \log_2 1 + 2 \log_2 2 + \ldots\ldots\ldots + 9 \log_2 9 + 3 \log_2 3] \quad ,$$

$$s_{jm} = 372.38710/125 \quad ,$$

$$s_{jm} = 2.97910 \quad .$$

We obtain the value for s_i from the n_i in the bottom marginal of Table I:

$$s_i = \frac{1}{125} [31 \log_2 31 + 30 \log_2 30 + 33 \log_2 33 + 31 \log_2 31] \quad ,$$

$$s_i = 620.83188/125 \quad ,$$

$$s_i = 4.96665 \quad .$$

The computation for s is based on the total number of measurements:

$$s = \log_2 125 = 6.96579 \quad .$$

It is evident that these calculations are performed very easily with a table of $n \log_2 n$. If he wishes, the reader may also make the computations with tables of $p \log_2 p$ like those prepared by Newman[8] and Dolansky.[3] The use of $p \log_2 p$ tables for analyzing discrete data is not recommended, however, because it leads to rounding errors that the table of $n \log_2 n$ avoids. The complete set of s-terms in the experiment on frequency judgment worked out as follows:

$s_{ijm} = 1.45211$	$s_i = 4.96665$
$s_{ij} = 2.91389$	$s_j = 4.79269$
$s_{im} = 2.99246$	$s_m = 4.93380$
$s_{jm} = 2.97910$	$s = 6.96579$

In Sec. V it was shown that response information H'(y) can be analyzed into components

$$H'(y) = H'_{uv}(y) + T'(u;y) + T'(v;y) + A'(uvy) \quad . \qquad (17)$$

Since $H'(y) = s - s_m$, we see that H'(y) = 2.03199 bits. If the subject had used the four response keys equally often, this figure would have been at most 2 bits. The extra information shows that the subject sometimes did not respond. This can be verified from the right-hand marginals in Tables I and II. The rest of the quantities in Eq. (17) are easily computed from s-terms. For example, $H'_{uv}(y)$ is computed from $s_{ij} - s_{ijm}$. We see that $H'_{uv}(y)$ is 1.46178 bits. This is the part of the response information that is not accounted for by either the auditory stimuli or the presponses. Consequently, 1.46178/2.03199 or 72 per cent of the response information is unanalyzed error. Some 28 per cent of the response information must therefore be due to associations between the subject's responses and the two predicting variables.

If we consider the association between auditory stimuli (u) and responses (y), we have

$$T'(u;y) = s - s_i - s_m + s_{im} \quad ,$$

$$T'(u;y) = 0.05780 \quad .$$

Thus only 0.058 bits are transmitted from the frequency stimuli, accounting for less than 3 per cent of the response information. This is not surprising because the signal-to-noise ratio was set near the masked threshold, and the stimuli were difficult to hear.

If we consider the association between presponses (v) and current responses (y), we find a little more transmitted information:

$$T'(v;y) = s - s_j - s_m + s_{jm} \quad ,$$
$$T'(v;y) = 0.21840 \quad .$$

This value of 0.218 bits transmitted amounts to some 11 per cent of the response information.

The last element in Eq. (17) is the stimulus × response × presponse interaction A'(uvy) This is computed from

$$A'(uvy) = -s + s_i + s_j + s_m - s_{ij} - s_{im} - s_{jm} + s_{ijm} \quad ,$$
$$A'(uvy) = 0.29401 \quad .$$

We see that about 14 per cent of the response information is due to the interaction. Knowledge of the interaction also permits us to hold one of the inputs constant while measuring transmission from the other input. For example, the transmission from stimuli to responses with presponses held constant is:

$$\begin{aligned} T'_v(u;y) &= s_j - s_{ij} - s_{jm} + s_{ijm} \\ &= T'(u;y) + A'(uvy) \\ &= 0.35181 \quad . \end{aligned}$$

Our calculations for the parts of the response information that we can analyze with the three-dimensional model lead to weights of approximately 3, 11 and 14 per cent for stimuli, presponses and interaction respectively. These figures sum to 28 per cent, the amount of transmitted information we predicted from the size of the noise term. We can also obtain this total weight directly by computing the information transmitted from both inputs together. We have

$$T'(u,v;y) = s - s_m - s_{ij} + s_{ijm} \quad ,$$
$$T'(u,v;y) = 0.57021 \quad .$$

If we now divide this three-dimensional transmitted information by the response information, we get back our figure of 28 per cent.

There are several points worth noting about our application of information theory to this experiment. The first is that the analysis is additive. The component measures of association, plus the measure of error (or noise), sum to the response information. Furthermore, the analysis is exact. No approximations are involved. The process is very similar to the partition of a sum of squares in analysis of variance. As a matter of fact, a notation can be worked out in analysis of variance that is exactly parallel to the s-notation in multivariate information transmission.[4]

The second point is that information transmission is made to order for contingency tables. Measures of transmitted information are zero when variables are independent in the contingency-sense (as opposed to the restriction to linear independence in analysis of variance). In

addition, the analysis is designed for frequency data in discrete categories, while methods based on analysis of variance are not. No assumptions about linearity are introduced in multivariate information transmission. Furthermore, when statistical tests are developed in a later section, it will be shown that these tests are distribution-free in the sense that they are extensions of the familiar chi-square test of independence.

The measure of amount of information transmitted also has certain inherent advantages. Garner and Hake[2] and Miller[5] have pointed out that the amount of information transmitted is approximately the logarithm of the number of perfectly discriminated input classes. In experiments on discrimination like the one that we have discussed, the measure provides an immediate picture of the subject's discriminative ability. Miller has also discussed applications of this property in mental testing and in the general theory of measurement.

VII. INDEPENDENCE IN THREE-DIMENSIONAL TRANSMISSION

It is evident from the definition of transmitted information that $T'(u,v;y) = 0$ when the output is independent of the joint input, i.e:, when

$$n_{ijm} = \frac{n_{ij} \cdot n_m}{n} \quad . \tag{20}$$

With this kind of independence, we can show that

$$s_{ijm} = s_{ij} + s_m - s \quad .$$

This expression for s_{ijm} may be substituted into Eq. (5) to confirm the fact that $T'(u,v;y) = 0$.

Now suppose that $T'(u,v;y) > 0$, but that v and y are independent, that is to say,

$$n_{jm} = \frac{n_j \cdot n_m}{n} \quad . \tag{21}$$

This leads to

$$s_{jm} = s_j + s_m - s \quad .$$

If we substitute for s_{jm} in Eq. (9), we find that $T'(v;y) = 0$. Equation (21) does not provide a unique condition for independence between v and y. To show this, let us pick some value of u and study the v-to-y transmission at that value of u. We now require that

$$n_{ijm} = \frac{n_{ij} \cdot n_{im}}{n_i} \quad . \tag{22}$$

If we have Eq. (22) for all i, we must have

$$s_{ijm} = s_{ij} + s_{im} - s_i \quad ,$$

and it follows, from substitution in Eq. (10), that $T'_u(v;y) = 0$. This is the situation in which v and y are independent, provided that u is held constant. It is an interesting case because we can show from Eq. (14) that if this kind of independence happens,

$$A'(uvy) = - T'(v;y) \quad .$$

The sign of T'(v;y) must be positive or zero so that – T'(v;y) must be negative or zero. Consequently, A'(uvy) can be negative. We see that negative interaction information is produced when the information transmitted between a pair of variables is due to a regression on a third variable. Holding the interacting variable constant causes the transmitted information to disappear.

If we have the independence defined by Eq. (21), we may not necessarily have the independence defined by Eq. (22). Let us suppose that we have both, i.e., that we have

$$s_{jm} = s_j + s_m - s \quad ,$$
$$s_{ijm} = s_{ij} + s_{im} - s_i \quad .$$

Now we substitute for s_{jm} and s_{ijm} in Eq. (8).

$$T'_v(u;y) = s_j - s_{ij} - s_{jm} + s_{ijm} \quad ,$$
$$T'_v(u;y) = s_j - s_{ij} - s_j - s_m + s + s_{ij} + s_{im} - s_i \quad ,$$
$$T'_v(u;y) = s - s_i - s_m + s_{im} \quad ,$$
$$T'_v(u;y) = T'(u;y) \quad .$$

Both kinds of independence, Eqs. (21) and (22), together mean that v is not involved in transmission between u and y. When this happens, we do not have three-dimensional transmission, since u is the only input variable.* As might be expected, both kinds of independence can be generated from a single restriction on the data, namely,

$$n_{ijm} = \frac{n_{im}}{V} \quad ,$$

where V is the number of classes in v.

We have studied the case where v is independent of y. We could have had u independent of y, or u independent of v. The results are analogous to those we have presented.

VIII. CORRELATED SOURCES OF INFORMATION

Three-dimensional transmitted information, T'(u,v;y), accounts for only part of the total amount of association in a three-dimensional contingency table. It does not exhaust all the association in the table because it neglects the association between the inputs. When this association is considered, i.e., when all the relations in the contingency table are represented, we are led to an equation that is very useful for generating the components of multivariate transmission. Consider

$$C'(u,v,y) = H'(u) + H'(v) + H'(y) - H'(u,v,y) \quad . \tag{23}$$

If we add and subtract H'(u,v), we obtain

$$C'(u,v,y) = T'(u;v) + T'(u,v;y) \quad ,$$
$$C'(u,v,y) = T'(u;v) + T'(u;y) + T'(v;y) + A'(uvy) \quad . \tag{24}$$

*Provided that no information is transmitted between u and v.

We see that C'(u,v,y) generates all possible components of the three correlated information sources u,v and y.

IX. FOUR-DIMENSIONAL TRANSMITTED INFORMATION

It will be instructive to extend our measures one step further, i.e., to transmitted information with three input variables, since from that point results can be generalized easily to an N-dimensional input. For simplicity, we shall restrict our development to the case of a channel with a multivariate input and univariate output. The more general case with N inputs and M outputs does not present any special problems, and can be constructed with no difficulty once the rules become clear.

Let us add a new variable w to the bivariate input u,v. The joint input is now u,v,w. We suppose that w sends signals h = 1,2,3, ... W. This gives us four sources of information u,v,w, and y. We can proceed to define a four-way interaction information A'(uvwy) as follows:

$$A'(uvwy) = A'_w(uvy) - A'(uvy) \quad .$$

We have already defined A'(uvy). The definition of $A'_w(uvy)$ will be similar, except that the subscript w indicates that A'(uvy) is to be averaged over w. As we have already noted, this is accomplished by adding the subscript h to each of the s-terms that make up A'(uvy). Consequently,

$$A'_w(uvy) = -s_h + s_{hi} + s_{hj} + s_{hm} - s_{hij} - s_{him} - s_{hjm} + s_{hijm} \quad . \tag{25}$$

It is readily shown that A'(uvwy) is symmetrical in the sense that it does not matter which variable is chosen for averaging, i.e.,

$$\begin{aligned} A'(uvwy) &= A'_u(vwy) - A'(vwy) \quad , \\ &= A'_v(uwy) - A'(uwy) \quad , \\ &= A'_w(uvy) - A'(uvy) \quad , \\ &= A'_y(uvw) - A'(uvw) \quad . \end{aligned} \tag{26}$$

We see that A'(uvwy) is the amount of information gained (or lost) in transmission by controlling a fourth variable when any three of the variables are already known.

If we examine all possible associations in a four-dimensional contingency table, we obtain

$$\begin{aligned} C'(u,v,w,y) = {} & T'(u;v) + T'(u;w) + T'(u;y) + T'(v;w) + T'(v;y) + T'(w;y) \\ & + A'(uvw) + A'(uvy) + A'(uwy) + A'(vwy) + A'(uvwy) \quad , \end{aligned} \tag{27}$$

where

$$C'(u,v,w,y) = H'(u) + H'(v) + H'(w) + H'(y) - H'(u,v,w,y) \quad .$$

Equation (27) can be proved by expanding both sides in s-notation. It turns out that, in the general case, C'(u,v,w, ... y) is expanded by writing down T-terms for all possible pairs of variables, and A-terms for all possible combinations of three, four variables and so on.

Four-dimensional transmitted information from u,v,w, to y, i.e., T'(u,v,w;y), can be written as follows:

$$T'(u,v,w;y) = H'(y) + H'(u,v,w) - H'(u,v,w,y) \quad . \tag{28}$$

The same arguments are used to justify Eq. (28) as were used in the case of Eq. (4) in three-dimensional transmission. To find the components of $T'(u,v,w;y)$, we note that

$$T'(u,v,w;y) = C'(u,v,w,y) - C'(u,v,w) \quad . \tag{29}$$

This means that $T'(u,v,w;y)$ contains all the components of $C'(u,v,w,y)$ except the correlations among the inputs. Consequently, the components of $T'(u,v,w;y)$ are

$$T'(u,v,w;y) = T'(u;y) + T'(v;y) + T(w;y) + A'(uvy) + A'(uwy) + A'(vwy) + A'(uvwy) \quad . \tag{30}$$

The components of $T'(u,v,w;y)$ are shown in schematic form in Fig. 2.

If it happens that

$$n_{hijm} = n_{ijm}/W \quad ,$$

where W is the number of classes in w, all the components of $C'(u,v,w,y)$ that are functions of w drop out, and $C'(u,v,w,y) = C'(u,v,y)$. In similar fashion, $C'(u,v,y)$ can be reduced to $C'(u,y)$. This is precisely what we did in the analysis of independence in three-dimensional transmitted information. Since $C'(u,y) = T'(u;y)$, we see that all cases of transmission with multivariate inputs can be related to the bivariate case.

Fig. 2. Schematic diagram of the components of four-dimensional transmitted information with three transmitters and a single receiver.

With three inputs controlled, we are ready to extend the analysis of response information in Sec. V a step further. We have

$$H'(y) = H'_{uvw}(y) + T'(u,v,w;y) \quad . \tag{31}$$

Equation (31) says that we can measure the effects in response information due to the three inputs This is evident from the fact that Eq. (30) tells us how to expand $T'(u,v,w;y)$ in its components. In addition, we know that

$$H'_{uv}(y) = H'_{uvw}(y) + T'_{uv}(w;y) \quad , \tag{32}$$

where

$$T'_{uv}(w;y) = T'(w;y) + A'(uwy) + A'(vwy) + A'(uvwy) \quad . \tag{33}$$

We see that controlling w in addition to u and v enables us to rescue the information transmitted between w and y from the noise, and to replace $H'_{uv}(y)$ with a better estimate of noise information namely, $H'_{uvw}(y)$.

The transition to N-dimensional input is now evident. In general, we have

$$H'(y) = H'_{uvw\ \ldots\ z}(y) + T'(u,v,w,\ \ldots,z;y) \quad . \tag{34}$$

The N + 1 dimensional transmitted information T'(u, v, w, . . . , z; y) can then be expanded in its components in the manner that we have described.

X. ASYMPTOTIC DISTRIBUTIONS

Miller and Madow[6] have shown that sample information is related to the likelihood ratio. Following Miller and Madow, we can show that the large sample distribution of the likelihood ratio may be used to find approximate distributions for the quantities involved in multivariate transmission.

Consider, for example, three-dimensional sample transmitted information T'(u, v; y). We can test the hypothesis that T(u, v; y) is equal to zero. This is equivalent to the hypothesis that

$$p(i,j,m) = p(i,j) \cdot p(m) \quad , \tag{35}$$

since T(u, v; y) is zero when input and output are independent. This hypothesis leads to the likelihood ratio (cf. Ref. 7),

$$\lambda = \frac{n^{-2n} \prod_{i,j} (n_{ij})^{n_{ij}} \prod_{m} (n_m)^{n_m}}{n^{-n} \prod_{i,j,m} (n_{ijm})^{n_{ijm}}} \tag{36}$$

If we take logs, we obtain

$$\frac{-2 \log_e \lambda}{1.3863\, n} = s - s_m - s_{ij} + s_{ijm} \quad ,$$

$$-2 \log_e \lambda = 1.3863\, n\, T'(u, v; y) \quad . \tag{37}$$

For large samples, $-2 \log_e \lambda$ has approximately a χ^2 distribution with (UV – 1) (Y – 1) degrees of freedom when the null hypothesis of Eq. (35) is true. Thus 1.3863 n T'(u, v; y) is distributed approximately like χ^2 if T(u, v; y) is equal to zero.

A more important problem involves testing suspected information sources. Suppose in our three-dimensional example, we assume that

$$p(i,j,m) = p(i) \cdot p(j) \cdot p(m) \quad . \tag{38}$$

This hypothesis leads to the likelihood ratio for complete independence in a three-dimensional contingency table,

$$\lambda = \frac{n^{-3n} \prod_{i} (n_i)^{n_i} \prod_{j} (n_j)^{n_j} \prod_{m} (n_m)^{n_m}}{n^{-n} \prod_{i,j,m} (n_{ijm})^{n_{ijm}}} \tag{39}$$

After we take logs, we find that

$$\frac{-2 \log_e \lambda}{1.3863\ n} = 3s - s_i - s_j - s_m - s + s_{ijm}$$
$$= H'(u) + H'(v) + H'(y) - H'(u,v,y)$$
$$-2 \log_e \lambda = 1.3863\ n\ C'(u,v,y) \quad .$$

For large samples $-2 \log_e \lambda$ has approximately a χ^2 distribution with $(UVY-1) - (U-1) - (V-1) - (Y-1)$ degrees of freedom when the null hypothesis is true.

We also know that

$$C'(u,v,y) = T'(u;y) + T'(v;y) + T'_y(u;v) \quad . \tag{41}$$

The likelihood ratio can be used to show that $1.3863\ n\ T'(u;y)$ and $1.3863\ n\ T'(v;y)$ are asymptotically distributed like χ^2 with $(U-1)(Y-1)$ degrees of freedom and $(V-1)(Y-1)$ degrees of freedom, respectively, if $T(u;y)$ and $T(v;y)$ are zero. To find the asymptotic distribution of $T'_y(u;v)$, we make the following hypothesis:

$$p(i,j,m) = p(i,m) \cdot p_m(j), \tag{42}$$

where $p_m(j)$ is the conditional probability of j given m.

Now we have the ratio

$$\lambda = \frac{n^{-n} \prod_{i,m} (n_{im})^{n_{im}} \prod_{j,m} \left(\frac{n_{jm}}{n_m}\right)^{n_{jm}}}{n^{-n} \prod_{i,j,m} (n_{ijm})^{n_{ijm}}} \quad , \tag{43}$$

$$\frac{-2 \log_e \lambda}{1.3863\ n} = s_m - s_{im} - s_{jm} + s_{ijm} \quad ,$$

$$-2 \log_e \lambda = 1.3863\ n\ T'_y(u;v) \quad . \tag{44}$$

In this case, $-2 \log_e \lambda$ has $Y(U-1)(V-1)$ degrees of freedom. In view of Eq. (41), we can write

$$1.3863\ n\ C'(u,v,y) = 1.3863\ n\ [T'(u;y) + T'(v;y) + T'_y(u;v)] \quad . \tag{45}$$

The quantities on the right side of Eq. (45) have degrees of freedom that sum to $(UVY-U-V-Y+2)$ Since this is the same number of degrees of freedom as on the left hand side of Eq. (45), the quantities on the right side of Eq. (45) are asymptotically independent, if the null hypothesis

$$p(i,j,m) = p(i) \cdot p(j) \cdot p(m)$$

is true.

This means that, as an approximation, we can test $T'(u;y)$, $T'(v;y)$ and $T'_y(u;v)$ simultaneously for significance under the null hypothesis we have stated. The test is very similar to an analysis of variance. We can see the similarity by applying the test to the data from our example in Sec. VI. The significance tests will be made on the quantities in Eq. (45). To do this, we need to compute $C'(uvy)$ and $T'_y(u;v)$, since these terms were not discussed in Sec. VI. First, we note that $C'(uvy)$ is the total amount of association in the stimulus × response × presponse table. We have

$$C'(uvy) = 2s + s_{ijm} - s_i - s_j - s_m \quad ,$$

$$C'(uvy) = 0.69055 \quad .$$

We also need $T'_y(u;v)$, the information transmitted from presponses to stimuli with responses held constant. This measures how successfully the presponses predict the auditory stimuli. Since stimuli were chosen at random, we do not expect much transmitted information here. The computation goes as follows:

$$T'_y(u;v) = s_m - s_{im} - s_{jm} + s_{ijm} \quad ,$$

$$= T'(u;v) + A'(uvy) \quad ,$$

$$= 0.41435 \quad .$$

We may now put our computed values for C'(uvy), T'(u;y), T'(v;y) and $T'_y(u;v)$ into Eq. (45) and perform the χ^2 tests. The results are summarized in Table III. We have not attempted to calculate the significance level of C'(uvy) because we do not have enough data to sustain the 88 degrees of freedom. The same criticism can probably be leveled at our test for $T'_y(u;v)$. In any case, Table III shows that the only significant effect in the experiment is the presponse-response association.

TABLE III

TABLE OF TRANSMITTED INFORMATION

Transmission	Component	$-2 \log_e \lambda$	Degrees of Freedom	P
Stimulus-Response	T'(u;y)	10.016	12	>.50
Presponse-Response	T'(v;y)	37.844	16	<.01
Presponse-Stimulus	$T'_y(u;v)$	71.802	60	=.14
Total	C'(u,v,y)	119.664	88	

One interesting fact that the analysis brings out clearly is that we cannot decide whether an amount of transmitted information is big or small without knowing its degrees of freedom. In our example we find the $T'_y(u;v) = 0.414$ bits, while $T'(v;y) = 0.218$ bits. Yet T'(v;y) is significant and $T'_y(u;v)$ is not. The reason lies in the difference in degrees of freedom. Miller and Madow[6] have discussed the amount of statistical bias in information measures due to degrees of freedom, and have suggested corrections.

In Table III, we tested $T'_y(u;v)$, the association between presponses and stimuli with responses held constant. This association is broken down still further in Table IV. No probability is estimated in Table IV for the interaction term A'(uvy) because its asymptotic distribution is not chi-square. All A-terms are distributed like the difference of two variables, each of which has the chi-square distribution. The distribution of this difference is evidently not chi-square because the difference can be negative. Its density function has been derived by Pearson,

Stouffer, and David[9], but the writer has been unable to find a table of the integral. In some cases the problem can be circumvented by combining A-terms with T-terms to make new T-terms. [See, for example, Eq. (33).] However, in other cases, the interactions are genuinely interesting in their own right, and should be tested directly. These cases can be treated when adequate tables become available.

TABLE IV

TABLE OF TRANSMITTED INFORMATION

Transmission	Component	$-2 \log_e \lambda$	Degrees of Freedom	P
Presponse-Stimulus	T'(u;v)	20.853	12	>.05
Interaction	A'(uvy)	50.948		**
Total	$T'_y(u;v)$	71.802	60	=.14

**Probability not estimated.

REFERENCES

1. R. M. Fano, "The Transmission of Information – II," Technical Report No. 149, Research Laboratory of Electronics, M. I. T. (6 February 1950).
2. W. R. Garner and H. W. Hake, Psychol. Rev. 58, 446-459 (1951).
3. L. Dolansky, "Table of p log p," Technical Report No. 227, Research Laboratory of Electronics, M. I. T. (2 January 1952).
4. W. J. McGill, "Multivariate Transmission of Information and its Relation to Analysis of Variance," Report No. 32, Human Factors Operations Research Laboratories, M. I. T. (May 1953).
5. G. A. Miller, Amer. Psychologist 8, 3-11 (1953).
6. G. A. Miller and W. J. Madow, "Information Measurement for the Multinomial Distribution" (in preparation).
7. A. M. Mood, *Introduction to the Theory of Statistics* (McGraw-Hill Book Co., Inc., New York, 1950).
8. E. B. Newman, Amer. Jour. Psychol. 64, 252-262 (1951).
9. K. Pearson, S. A. Stouffer and F. N. David, Biometrika 24, 293-350 (1932).
10. C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949).
11. J. E. Keith Smith, "Multivariate Attribute Analysis" (in preparation).
12. F. L. Stumpers, "A Bibliography of Information Theory," Technical Report, Research Laboratory of Electronics, M. I. T. (2 February 1953).

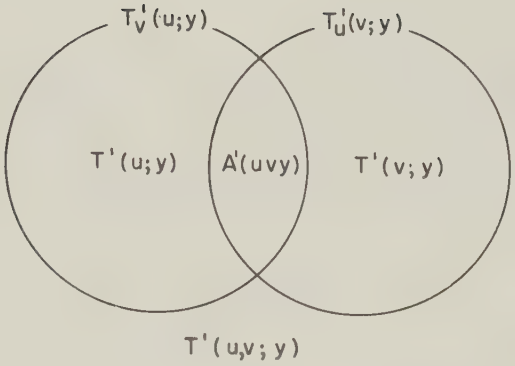


Fig. 1 - Schematic diagram of the components of three-dimensional transmitted information. The diagram shows that three-dimensional transmission can be analyzed into a pair of bivariate transmissions plus an interaction term. The meanings of the symbols are explained in the text.

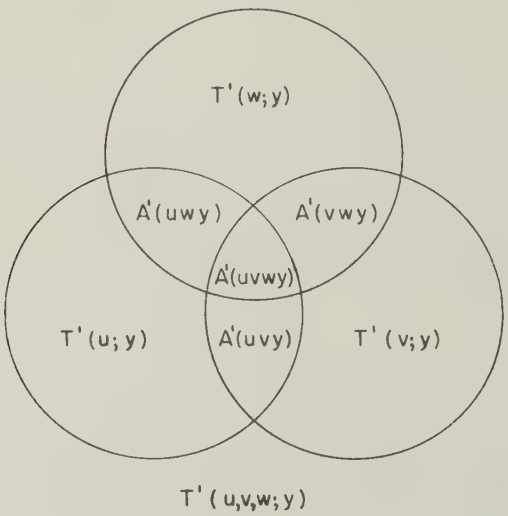


Fig. 2 - Schematic diagram of the components of four-dimensional transmitted information with three transmitters and a single receiver.

CHOICE AND CODING IN INFORMATION RETRIEVAL SYSTEMS

Calvin N. Mooers
Zator Company
Boston, Mass.

Introduction

Information retrieval machines are devices for indexing and selecting information in a library. The operation of these machines is based upon some sort of an arbitrary code system, in terms of which the machines' operations are defined. Because they use coding systems to deal with information, such machines have much in common with devices for point-to-point signalling, and in particular, with multiplex transmission systems. One is thus led to inquire whether--or in what manner--the formalism of communication theory as developed for signalling can be applied to machines for information retrieval. This paper presents several results from such an inquiry.

In general, it can be said that the methodology and approach of communication theory has been helpful in building a theory of information retrieval systems. Three topics are discussed in this paper. 1) The retrieval system analogue to $H = -\sum p_i \log p_i$, the measure of the output of a source, is developed. 2) The retrieval system analogues to synchronous and asynchronous multiplex types of coding are described and channel capacities are discussed. For the latter type of coding, called "superimposed," a new limit on channel capacity of $\log_e 2$ bits per site is given. 3) Selection errors due to coding are discussed. It is shown that the frequency of errors can be made arbitrarily small for the superimposed type of coding, in analogue to the result for signalling.

The Retrieval System Model

While there are a variety of retrieval systems, such as those based upon decimal classifications or upon alphabetical indexing, this paper is restricted to those systems in which a machine serially scans the marks on each tally from a battery of tallies to make the selection of the desired documents.[1] For instance, the machine may be one which scans the punches in a pack of Hollerith cards, making certain selections as it does so. Each document in the library is represented by a tally, and marks and blanks on the tally carry in digital form indications of the subject matter in the document.

There are a variety of opinions and proposals on how to deal with the semantic aspects of the subject matter of a document to prepare it for the digital coding. Unlike the situation in communication theory, the semantic problem cannot be dodged in information retrieval. The method to be described here is thus not the only one. However, it has the advantage of having met the test of wide usage, and of being simple. In essence, there is a restricted set or repertory of semantic units called "descriptors."[2,3] Each descriptor stands for a pre-assigned scope of meaning. For each document, a subset is formed from those descriptors whose scope of meaning touches upon the semantic content of the document. Other than their being grouped in a subset, no other interrelationship is made between the descriptors of a subset. For document D_i, we shall denote its descriptor subset by $D_i(S)$. A selection of useful documents is prescribed in terms of one, two, or more descriptors conjointly. If the descriptors S_a, S_b, and S_c are in the selection prescribing subset R(S), we require that the selector machine shall segregate tallies of documents D_i for which all of the descriptors in the subset R(S) are contained within the subset $D_i(S)$.

It will help to have before us some numbers illustrative of the problem. The number of documents in a typical collection, and thus the number B of tallies in a battery, may be in the order of 10,000. On a tally the number of sites F available for code marks and blanks is in the order of 200, though in different operating systems it may range from 40 to 500. In practice it has often been found that an average of approximately 8 or 10 descriptors are used in the characterization of any document such as a technical report, journal article, or the like. In practice it is often possible to set an upper limit k_b on the number of descriptors in the subsets $D_i(S)$, where k_b is almost never exceeded. At the other end, it is also often possible to set a lower limit k_a on the number of descriptors in the selection prescribing subsets R(S), such that the number of these descriptors almost never is less than k_a. The existence of such "limits" in practice is very important.

The Measure Of Required Choice Per Tally, J

In communication theory the quantity $H = -\sum p_i \log p_i$ is a measure of variety in the output of a source. Described in another way, H is in the nature of a weighted average of the minimal number of binary digits (bits) required to express the choice between the possible messages of a source, the messages being independent, and p_i being the a priori probability of the i^{th} message. The measure of choice H is important because it directly determines the least amount of digital "content" that a signalling channel may transmit in order that a receiving terminal may reproduce the output of the source.[4]

As a matter of terminology, and to prevent confusion, we shall refrain here from identifying H with "amount of information." H is a measure of digital representations and their frequencies--abstracted from semantic information. Information retrieval systems, on the other hand, must deal with very real problems in semantic information on at least two levels: the document text, and the descriptors.

We shall now inquire into the retrieval system analogue of the H of communication theory. A heuristic rather a rigorous argument is followed here as well as later, such an argument being appropriate to the present state of the retrieval art. A retrieval selection is prescribed in terms of a descriptor subset R(S) taken from a set of descriptors numbering V. The result of the retrieval operation is the selection of a subset of tallies from the B tallies in the battery. The selected subset may have no tally, one tally, or many tallies--with all cases being useful. Fortunately we do not have to deal with the very general problem of mapping all subsets into subsets. The nature of the information in the documents narrows the problem. We specifically note that the operation of choice during machine selection is upon the battery of tallies numbering B, and not upon the descriptor set numbering V. There is (in the first approximation) no reason to expect some tallies to be chosen more frequently than the others. Thus, the specification for choice of any one tally from the set of B tallies requires a minimum of $\log_2 B$ bits to express the choice. Furthermore, the specification is made in terms of at least k_a descriptors acting conjointly. Presuming also (as a first approximation) that the various descriptors are used with equal frequency, each descriptor need carry more than $(1/k_a)\log_2 B$ bits of choice.

Each tally, in its code of marks and blanks, must indicate a capability for selection according to the descriptors $D_i(S)$ which number k_i. The number F of sites available for marking on a tally is fixed (according to the model under consideration). We cannot do as in communication theory and consider merely the average number of bits required. We must consider instead the case of the maximum number of bits, and be sure that this case is under control. The maximum is the case of a tally having k_b descriptors. Therefore, in the model given, a tally must have a capacity for indicating a choice of

$$J = (k_b/k_a)\log_2 B \quad \text{bits per tally.} \qquad (1)$$

The quantity J plays the part in information retrieval which is analogous to the part played by H in communication theory. H sets the minimal channel capacity compatible with the accurate reproduction of the output of the source. J sets the minimal tally capacity compatible with the accurate retrieval selection upon the set numbering B.

We specifically note that for a less restricted model, the quantity J can be somewhat smaller, and in fact various weighted means become possible. One such is $J = \sum_i (k_i/k_a)\log_2 B$ in the case the effective F is permitted to vary in certain ways. This is no surprise, since H in communication theory can also become smaller when account is taken of any non-independence of the output probabilites of the source. Aside from mentioning that J can be refined in various directions we shall not pursue the matter further. On a practical level, the value in (1) is adequate for specifying the problem.

As a numerical example, we take the illustrative case given previously. For B = 10,000 and with the typical values of $k_a = 2$ and $k_b = 15$, the required choice per tally is J = 100 bits.

The Two Kinds Of Coding Systems

There are two basically different kinds of coding now being used in information retrieval systems. For convenience, they are distinguished here by the terms "binary coding" and "superimposed coding." The latter, in the form of superimposed random coding, is also known under the trade name "Zatocoding."[5] Coding in retrieval systems is closely analogous to coding in multiplexed signalling systems. On each tally, which may be compared to a time segment of a signalling channel, each descriptor must operate effectively and independently of the others. Retrieval system "binary coding" is very much like a time division pulse code multiplex signalling system. Superimposed coding is very much like an asynchronous time division multiplex signalling system--although its exact signalling parallel has not yet appeared in practice. Related systems have been mentioned.[5,7] The salient features of the binary and superimposed coding methods are as follows:[8]

Binary	Superimposed
The F sites of each tally are partitioned into groups of N sites each, with each of the F/N groups able to take one descriptor code pattern.	The F sites are not partitioned; the descriptor code patterns rely upon statistical randomness to insure separation.
A descriptor pattern is a binary numeral consisting of marks and blanks with a total of N digits.	A descriptor pattern consists of marks only, with the N marks of a pattern being distributed over the F sites.

Binary	Superimposed
Descriptor patterns are added to a tally by putting the new pattern into any unoccupied group of sites.	Descriptor patterns are added to a tally by Boolean addition of the new pattern to the marks already on the tally, i.e., by pattern superimposition. The only tally sites left unmarked are those not having a mark from any descriptor code.
Tally selection according to a descriptor pattern requires that in some tally group of sites there is a pattern whose marks and blanks agree completely with the selecting descriptor pattern.	Tally selection according to a descriptor pattern requires that there is a mark in every tally site corresponding to a mark in the selective pattern, and other marks and blanks in the tally make no difference. The pattern of selecting marks is included within the pattern of tally marks.
Tally selection according to several conjoint descriptors requires that the pattern for each and every selecting descriptor must be found somewhere in the selected tally.	Tally selection according to several conjoint descriptors requires that the Boolean sum of all the selecting descriptor patterns must be included within the total tally pattern of marks.
Changing a mark to a blank on a tally, or vice versa, will completely change a code, giving it the indication of another descriptor.	Changing a blank to a mark on a tally does not preclude selection, and has little effect. Changing a mark to a blank may prevent several descriptors from selecting a tally.
Descriptor codes are non-interfering, being in separate partitions. There are no other sources of digital noise.	Superimposing the codes does cause a kind of interference by changing blanks to marks. Desired tallies are never excluded by the interference. The noise does result in the spurious appearance of a few extra tallies.
The tally is completely filled when there is a descriptor pattern in each partition.	The tally is optimally used when the density of marks approaches 50 per cent.
Selection according to k descriptors requires the making of the equivalent of k times F/N separate pattern-matching attempts for each tally scanned.	Selection according to any number of descriptors requires only one pattern-matching attempt per tally.

Coding according to somewhat inefficient versions of the "binary" method has been used widely for a very long time in information retrieval machines.[1] The "superimposed" method with random patterns is more recent. A detailed discussion of it will be found elsewhere.[3,5,9,10] A striking difference between the two methods of coding is the difference in complexity of the selector required. The binary coding selector machines must try many different pattern matchings, and as a result are either slow or expensive. Machines for superimposed coding, having a simpler task, are fast and relatively simple. In payment for this considerable advantage, superimposed random coding requires that the number of tally sites F be 45 per cent greater for the same J.

Capacity In Bits Per Tally -- Binary Coding

A battery of tallies, quite as well as a signalling channel, has a measurable capacity for carrying the coded indications of choice. In the case of binary coding, the battery itself is inherently noiseless--presuming no malfunction of the selecting machine or other accidents. It is thus plausible to speculate that each tally has a capacity of one bit per site, or F bits per tally. Arguments substantiating this speculation are given here. A peculiarity typical of information retrieval systems arises. While the "channel" itself is essentially noiseless, efficient codings which are capable of exploiting approximately the full capability of each tally must necessarily give rise to a small but finite selection error. That is, for binary coding systems wherein the number of tally sites F becomes approximately equal to the quantity J, a type of error appears that may be called "code synonym error."

The error stems from the fact that $J = (k_b/k_a)\log_2 B$ takes no accout of the number V of the descriptors in the repertory set. This omission is quite correct, however, since tally choice only involves a choice from among the B tallies. It does not require a dependence upon V. An efficient coding allows only $N = (1/k_a)\log_2 B$ digits per descriptor. However, N digits can form only 2^N different binary patterns, and V may be larger than 2^N. (When V is not larger, there is no synonym error.) For example, in the illustrative case already described, with $B = 10{,}000$, $k_a = 2$, and $V = 500$, the quantity N equals 7 and 2^7 is only 128. As a result, with such an efficient coding system, there is a four-fold doubling up. Four descriptors must use the same code pattern. However, no actual retrieval selections are specified by only a single descriptor at a time because $k_a = 2$. Consequently, the probability of error due to code synonyms is in the order of $(4/500)^2$ per tally,

which is somewhat less than 1/10,000. We note that the code synonym error does not erroneously exclude any tallies that should be selected. Instead it errs (in the worst case of two descriptors) by permitting the appearance of unwanted tallies, with a probability of less than one unwanted tally for each selection upon the entire battery of B tallies. In practical retrieval systems, such an error is no handicap whatsoever. The fact that B = 10,000 and that the error per tally is approximately 1/10,000 is not coincidental. It is a consequence of the logic behind the particular definition of J, requiring a maximal power of choice of one tally from a set of B tallies for the minimum of k_a prescribing descriptors.

The ordinarily-used retrieval systems using binary coding are not designed for such coding efficiency. This is due to the uncritical belief that descriptors must be coded with unique patterns, i.e., that N must be sufficiently large so that 2^N is greater than V. [1] In our illustration, an efficient binary coding (with the four-fold synonym redundancy) with F = 200 available sites, allows $k_b = 200/7 = 28$. This permits a choice-indicating ability of J = 186 bits per tally. (Because there cannot be fractional digits in the codes, the round-off errors will usually make J somewhat smaller than F, even for efficient coding.) In comparison, by the more wasteful, but more usual, coding of the same tally, for V = 500 we have N = 9. As a result $k_b = 22$ and J = 144 bits of choice per tally, an appreciable loss in capacity. Such a loss in choice per tally due to poor coding is even greater for larger values of V.

Following Shannon, in defining the channel capacity as the maximum transmission rate that can be achieved by a set of codings, we define the maximum tally capacity as the maximum number of bits of choice that can be carried. It is plausible to conclude that when the maximum is taken over the binary codings, the tally has a maximum capacity of one bit per site, or F bits per tally.

Capacity In Bits Per Tally -- Superimposed Coding

With superimposed coding some unexpected things happen when one tries to compute the analogue of channel capacity. The coding is inherently noisy, thus there is equivocation. Only the marks have meaning. The marks of one pattern interfere with other patterns, changing blanks to marks.

The design of superimposed coding systems for specific problems has been treated elsewhere.[3,8] The conclusions only are presented here. Each descriptor is given a pattern of $N = (1/k_a)\log_2 B$ marks. (This presumes that the design calls for an occurence of no more than one extra selection on the average for every scanning of the battery.) The descriptor patterns are generated by some random or quasi-random process, thus allowing a statistical separation of the patterns to operate. (This is equivalent to multiplexing on the basis of statistical improbability of the channels interfering inordinately.[7]) Optimum usage of the tallies requires that approximately 50 per cent of the tally sites be marked. For densities appreciably beyond this, interference, shown by extra selections, sharply increases. The 50 per cent criterion is met (conservatively) by setting the maximum k_b at the value $k_b = (\log_e 2)F/N = 0.69F/N$.

Let us now put these results aside, starting instead from the beginning with Shannon's formalism of communication theory.[4,6] We shall be concerned with finding the maximum capacity of a tally under superimposed codings. Let each tally with F sites be coded with k descriptors of N marks per code pattern. Because we use random codes, the probability that any site will have a mark from the k patterns is $P_k = 1 - (1 - N/F)^k$. For any one tally, consider the joint probability p(i,j) that when an indication "i" is marked into a site by a descriptor the indication "j" will be found on the fully marked tally. The a priori probability of any one site being marked by a descriptor is N/F. If a tally site is thus intentionally marked (indication "1"), the tally will remain marked at the site and p(1,1) = N/F. No mark will change to a blank, so p(1,0) = 0. For sites not intentionally marked by the descriptor--but possibly marked by any of the k-1 others--we have $p(0,1) = (1-N/F)(P_{k-1})$ and $p(0,0) = (1-N/F)(1-P_{k-1})$. With these assumptions, the transmission rate or tally capacity R in bits per site can be computed according to Shannon's formulas. The details are found in the appendix.

For all k descriptors of the tally, the total capacity must be k times the capacity for one descriptor, or $R_t = kR$ bits per site. Maximizing R_t as a function of k (see appendix) leads directly to these conditions on k,N, and F:

$$kN/F = \log_e 2 = 0.69 \qquad (2)$$

which in turn is equivalent to:

$$P_k = \tfrac{1}{2} \qquad (3)$$

These are the conditions for optimal use of the tally, giving it maximal capacity. Under these conditions, and for N/F small (as it usually is) the average tally capacity per descriptor is N/F bits per site (see appendix). Thus the tally carries just N bits for each descriptor. The total tally capacity under these optimal conditions is

$$kRF = (F/N)(\log_e 2)(N/F)F = F\log_e 2 \text{ total bits per tally} \qquad (4)$$

or

$$\log_e 2 = 0.69 \text{ bits per site.} \qquad (5)$$

From this it is apparent that the total capacity of a tally with superimposed coding cannot exceed $\log_e 2$ bits per site. Similarly, in any asynchronous multiplex communication system based upon any analogous use of superimposed signalling patterns, the transmission rate cannot exceed 0.69 bits per pulse width. This conclusion seems to be a new result for this kind of coding. However, in a narrower sense, the same conclusion was reached by the author by another method earlier.[9,10] Also, a figure in a paper by White contains a hint of the conclusion although he states he was unable to determine the maximum efficiency.[6]

Selection Errors, Noise And Coding

As has been shown, binary coding for retrieval selection has no "noise" and the only errors are due to coding synonyms in the case of an efficient coding. Superimposed coding is inherently noisy. With regard to it, we can look for an analogue to Shannon's result for a noisy channel. He has shown that it is possible so to encode a message that the probability of error becomes arbitrarily small, if sufficient time delay is allowed. In particular, he has stated that the probability of error in a given message sequence is bounded by

$$P \leq 2^{-T(C-H)} \tag{6}$$

where T is the length of the sequence, P is the probability of error, C is the channel capacity, and H is the rate of generation of the source. (It is to be noted that, in general, codings which achieve this result have not been achieved in practice.)

For superimposed coding, when no more than optimal coding density of marks is used, the probability P of error per tally in selection is bounded by

$$P \leq 2^{-G} \tag{7}$$

where G is the total number of marks in the superimposed selection pattern from the selection-prescribing descriptors of the subset R(S). The result is simply explained. At optimal use, the density of marks averages $\frac{1}{2}$. An erroneous selection can occur only if there is a chance occurence of a mark in each of the G sites of the selective pattern. For each site the chance is $\frac{1}{2}$, and for G sites simultaneously, it is less than 2^{-G}. It is to be noted that superimposed codings do in practice achieve results of this order, though deviations from it may occur due to correlations between certain descriptors as they are used to describe the documents.

This last analogy is no mere formal accident. Looking to the communication side of the analogy, we see that the message contained in the sequence of length T can be expressed in TH bits. This leaves T(C-H) bits of redundancy or choice for use in case of equivocation to retrieve the correct message from other possible but non-wanted messages. Each of the T(C-H) bits gives an improvement in error by a factor of $\frac{1}{2}$.

On the information retrieval side of the analogy, the actual message is in the document, so there is no purpose to coding it on the tally. All of the G marks of the selective pattern are available for choice to overcome the equivocation between documents of the collection and to retrieve the desired documents. Each of the G marks decreases the error by a factor of $\frac{1}{2}$.

Concluding Remarks

We have seen that information retrieval systems are susceptible to treatment by communication theory at the coding and machine level, and that there are a number of analogies between retrieval systems and multiplex signalling systems. Historically, retrieval theory has been aided by communication theory. In the other direction, there is reason to believe that developments--both theoretical and practical--originally made with retrieval systems may be applicable to the development of signalling systems. For instance, some retrieval practice seems to be ahead of work in asynchronous multiplex signalling. For another thing, techniques for handling semantic information in retrieval--not discussed here--may be suggestive for further development in communication theory if and when such matters are undertaken.

Appendix

We shall verify for the case of superimposed random coding that the maximum tally capacity occurs when $kN/F = \log_e 2$ with N/F small, and that under these circumstances the number of bits per descriptor is N. Where $p = N/F$, $q = 1 - p$; $P = P_{k-1} = 1 - (1 - p)^{k-1}$, and $Q = 1 - P$, and following Shannon's formalism closely:

$$\begin{pmatrix} p(1,1) & p(1,0) \\ p(0,1) & p(0,0) \end{pmatrix} = \begin{pmatrix} p & 0 \\ qP & qQ \end{pmatrix} \tag{8}$$

$$H(x,y) = -\sum_{ij} p(i,j) \log p(i,j) = -p \log p - qP \log qP - qQ \log qQ \tag{9}$$

$$H(x) = -\sum_{ij} p(i,j) \log \sum_{j} p(i,j) = -p \log p - q \log q \tag{10}$$

$$H(y) = -\sum_{ij} p(i,j) \log \sum_{i} p(i,j) = -(p+qP) \log (p+qP) - qQ \log qQ \tag{11}$$

$$R = H(x) + H(y) - H(x,y) \tag{12}$$

$$= -q \log q + qP \log qP - (p+qP) \log (p+qP) \tag{13}$$

$$R_t = kR \tag{14}$$

Maximizing R_t as a function of k:

$$\frac{\partial R_t}{\partial k} = R + k \frac{\partial R}{\partial k} = 0 \tag{15}$$

Make use of:

$$\frac{\partial P}{\partial k} = (-1) \log_e (1-p)(1-p)^{k-1} \tag{16}$$

$$-kq \frac{\partial P}{\partial k} = (1-p)^k \log_e (1-p)^k \tag{17}$$

$$\frac{\partial R_t}{\partial k} = R + k \cdot \begin{pmatrix} -q \log (p+qP) \\ -q \\ +q \log qP \\ +q \end{pmatrix} \cdot \frac{\partial P}{\partial k} \tag{18}$$

$$= R - (1-p)^k \log_e (1-p)^k \log \frac{qP}{p+qP} \tag{19}$$

Set $(1-p)^k = \frac{1}{2}$ throughout. This is the same as $P_k = \frac{1}{2}$, or $kN/F = \log_e 2$. (20)

$$\frac{\partial R_t}{\partial k} = -(1-p)\log(1-p) - \left(\tfrac{1}{2}\right)\log\left(\tfrac{1}{2}\right) + \left(\tfrac{1}{2}-p\right)\log\left(\tfrac{1}{2}-p\right) \qquad (21)$$

$$-\left(\tfrac{1}{2}\right)\log_e\left(\tfrac{1}{2}\right)\log\frac{\frac{1}{2}-p}{\frac{1}{2}}$$

Expand for small p, term by term:

$$= -(\log_2 e)(-p) + \tfrac{1}{2} + \left[-\tfrac{1}{2} + (-1+\log_2 e)(-p)\right] + \left(\tfrac{1}{2}\right)(\log_e 2)(\log_2 e)(-2p) \qquad (22)$$

$$= \left\{p(\log_2 e) - p(\log_2 e) + p\right\} - p = 0 \qquad (23)$$

The first term is R, so:

$$R = p = N/F \qquad (24)$$

References

1 R. S. Casey & J. W. Perry, editors, "Punched Cards, Their Applications To Science and Industry," New York, Reinhold, 1951.

2 C. N. Mooers, "Scientific Information Retrieval Systems For Machine Operation--Case Studies In Design," Zator Technical Bulletin No. 66, April 1951.

3 C. N. Mooers, "Zatocoding Applied To Mechanical Organization Of Knowledge," American Documentation pp.20-32, January 1951.

4 C. E. Shannon, "A Mathematical Theory Of Communication," Bell System Technical Journal, pp. 379-423, July 1948; and pp.623-656, October 1948.

5 British Patent No. 681,902 to C. N. Mooers (U.S. application Sept. 17, 1947); U.S. patents pending.

6 W. D. White, "Theoretical Aspects of Asynchronous Multiplexing," Proc. I.R.E., pp. 270-275, v. 38, March 1950.

7 J. R. Pierce & A. L. Hopper, "Nonsynchronous Time Division With Holding And With Random Sampling", Proc. I.R.E., v. 40, September 1952.

8 C. N. Mooers, "Zatocoding For Punched Cards," Zator Technical Bulletin No. 30, 1950.

9 C. N. Mooers, "Putting Probability To Work In Coding Punched Cards," presented at 112th meeting of the American Chemical Society, New York, September 1947. Published as Zator Technical Bulletin No. 10, 1947.

10 C. N. Mooers, "Application Of Random Codes To The Gathering of Statistical Information," master's thesis, Mathematics Department, M.I.T., February 1948.

MODERN STATISTICAL APPROACHES TO RECEPTION IN COMMUNICATION THEORY

David Van Meter*
David Middleton **, †
Cruft Laboratory, Harvard University
Cambridge, Massachusetts

Summary

When reception in the theory of communication is recognized as a problem in statistical inference, system design and system analysis appear as the counterparts of designing and evaluating statistical tests. This paper discusses the optimum properties of designs based on statistical decision theory from the risk point of view, and from that of information theory. Connections between risk and information loss are established, which result in a unified theory of system design. This includes Minimax methods capable in principle of handling all degrees of a priori knowledge of signal and noise statistics, new methods for comparing actual and ideal systems for the same purpose, and new interpretations of previously used formulations as special cases of the more general theory. Both detection and extraction of signals in noise are considered, the former as a problem of testing statistical hypotheses and the latter as one of estimating parameters.

Formulation of the general reception problem as a decision operation is followed by a summary of statistical decision theory from the risk point of view, with some examples of Bayes and Minimax tests and optimum classes of decision rules. Applications to detection show the optimum nature of likelihood ratio receivers as a class, and indicate methods for defining the minimum detectable signal and for comparing system performance. As an illustration, curves of Bayes and Minimax risk are given for detection of a pulsed carrier in noise. Applications to extraction show the nature of optimum extraction and the rôles of the mean square error and maximum likelihood criteria from the more general point of view of risk theory. Conditions under which information loss is an extremum in detection and extraction are established, and information loss itself as a criterion of performance is compared with that of the risk measure.

1. Introduction and General Formulation

1.0. Introduction

The central rôle of the noise background in reception and the noise-like character of many types of signal has naturally led to the application of various statistical ideas to the solution of reception problems. Of chief practical interest are the twin problems of "best" or optimum methods of detecting the presence or absence of a signal in noise, and of extracting a signal from a noisy background. Each of these classes includes many variants, which depend mainly upon what is known about the signal and noise and what is chosen as the criterion of "best." Perhaps earliest in point of time is the view that the best extractor is a transducer which accepts the mixture of signal and noise at its input and produces an output which is as close to the desired output as possible, the "distance" between them being defined in some appropriate way, usually as the squared error averaged over an interval equal to the time available for observation of the output. Wiener [1.1] originally treated the problem in this way and found suitable linear filters for stationary inputs and semi-infinite observation periods. Results have since been obtained for more general classes of filters and nonstationary inputs by Booton [1.4], Davis [1.5], Singleton [1.2], Zadeh and Ragazzini [1.3] among others.

The first treatments of detection were concerned with finding linear predetection filters which maximized the peak signal to rms noise ratio at the detector input, and hence at the detection output,

---

*On leave from Pennsylvania State University, State College, Pennsylvania

**Support for a small portion of this work was provided by the Department of the Army, the Department of the Navy, and the Department of the Air Force, under Contract with the Massachusetts Institute of Technology.

† Now at 49 Lexington Avenue, Cambridge.

when a monotonic relation between input and output is maintained. This marked a departure from the idea of controlling system errors, upon which the previously mentioned criterion for optimum extraction was based, and corresponds to a less complete view of the detection problem. The familiar "matched filter" for white noise, treated by North [1.6] and by Van Vleck and Middleton [1.7] was followed by the filters for "colored" noise of Den Hartog and Muller [1.8], Dwork [1.9], George [1.10], and Urkowitz [1.11]. Zadeh and Ragazzini [1.3] have considered problems of physical realizability and finite observation time which arise in this connection. Certain classes of non-linear predetection filters have also been discussed by Zadeh [1.12]. A second approach, which usually requires a higher order of a priori information than those based on the simpler "distance" criteria, deals with detection and extraction as operations which are analogous to hypothesis testing and parameter estimation in the theory of statistical inference (see, for example, Middleton [1.13]). System design and analysis then appear as the counterparts of designing and evaluating statistical tests. Grenander [1.14] has shown how classical methods of inference, which assume discrete uncorrelated samples, may be extended to stochastic processes.

Siegert's introduction [1.15] of the Ideal Observer as a model of the human observer in the pulsed radar case was followed by the use of statistical criteria for actual system design by Hanse [1.16], Slattery, Reich and Swerling [1.17], and others. The betting curve, used by Siegert to define the minimum detectable signal in a special instance was next employed by Middleton [1.18] in a unified presentation of several criteria appropriate for detection, including the sequential test. A somewhat different approach, used by Woodward and Davies [1.19] [1.20], viewed the receiver as an information processing device, and found that a receiver which presents the a posteriori probability of the signal at its output (but which makes no "decision"), transmits the most information about the input signal.

It is clear that there is an element of arbitrariness in whatever criterion of "optimum" is chosen. The important thing is obviously to fit the system design to the constraints of the particular problem at hand, with special attention to the amount and kind of statistical knowledge available and the type of result wanted. On the other hand, it is equally important not to impose unnecessary restrictions on the solution in order to obtain an easy answer without knowing the additional complexity or cost involved and the increase in performance to be obtained by lifting these restrictions. If investigation shows that certain statistical knowledge about signal or noise, not ordinarily available, would result in substantial design improvement, then certainly the cost of obtaining such information must be a factor in system planning.

Systems designed to maximize signal-to-noise ratio, [1.6-1.12], or to present a posteriori probability distributions [1.19], [1.20] may certainly be justified as "optimum" from some point of view, usually that the rôle of the reception system is to assist in making some specified judgment about the signal. Our interest here, however, is in systems that themselves actually make optimum (or near optimum) judgments or decisions, as it seems to us that these are of greater potential importance in practice.

In designing a detection system, for example, one is necessarily interested in the effects of the various errors which may occur and in the minimum detectable signal for given error performance. In a binary (or single-alternative) system, when "yes-or-no" is the required decision there are two types of error [1.13,1.18]: a Type I error, of mistaking noise for signal; and a Type II error, of mistaking signal for noise. It may be that Type I errors (false alarms) are much more important than Type II errors (false rests), e.g., an expensive sequence of operations may be initiated by an alarm, and none by a rest. Consequently there must be provision in the design for different relative weightings of the two types of error. Furthermore, the occurrence of errors depends critically upon the a priori probabilities of the occurrence of signal and noise. Usually, reliable estimates of these a priori probabilities are not easy to obtain, but a requirement of high-grade performance in a given situation may warrant effort in this direction. Certainly their possible effect must be taken into account in the analysis.

It is clear that a criterion of maximum signal-to-noise ratio at the output of a predetection filter is not by itself adequate when errors are important, since no mechanism for making decisions as to the presence or absence of a signal is considered. Neither does the (s/n)-approach tell us the effects of omitting or not optimally using the available information, or how such information should be best processed. If such a device is added to the system, the errors can be studied, but then the over-all system almost always incorporates an unnecessary constraint in the method of handling the data. It is more reasonable, and certainly it comes far closer to answering the question of finding the "best" system for a given class of decisions, to assume at the outset that the system must make these decisions and then to design it accordingly, making best use of all available statistical data within the external constraints imposed by the nature of the application. If this optimum system is

too expensive to realize, then compromises can be readily made, and evaluated by comparison with the optimum. [An example of this is given in Sec. (3.5)].

The present paper, it is believed, gives in large part a new approach to problems of reception in communication theory, an approach which adapts certain techniques recently developed by statisticians, to the design of detection and extraction systems. These methods here employ risk and information loss as two possible criteria of performance. In the risk formulation a cost is preassigned to each possible error the system can make, and the risk is calculated as the expected value of the cost in view of the various error probabilities involved. The best system in this instance is defined as one that minimizes this risk. In the formulation based on information loss, the measure of information [2.4] is used to calculate system equivocation and the properties of systems that minimize this quantity are investigated. The risk formulation has the advantage that with preassigned costs, it brings into the open the element of arbitrariness inherent in optimum criteria, allowing its effects to be studied. Special cost assumptions lead to previously used criteria such as the Neyman-Pearson and Ideal Observers [1.18] in detection, and the least mean-squared-error criterion in extraction [1.1,1.12]. Minimax methods for the design of tests when the available statistical data are incomplete, and methods for comparing actual and ideal systems complete our present theory of optimum system structurization. These provide techniques which appear to fit more closely than earlier efforts, the actual conditions under which practical solutions to detection and extraction problems are required.

1.1. General Formulation

We begin by considering the principal elements of a typical decision situation at a level of generality which does not restrict us to any specific type of signal, noise, or decision, and which does not at this point involve special assumptions about the statistics of signal and noise. Figure 1 shows such a general formulation, with the pertinent notation. A decision $\underset{\sim}{\gamma}$ is to be made about a signal $\underset{\sim}{S}$ based on observations $\underset{\sim}{V}$ of the mixture of signal and noise $\underset{\sim}{V} = \underset{\sim}{S} \otimes \underset{\sim}{N}$. In general, signal and noise are functions of time, and observation is confined to a finite time interval (0,T). These observations may consist of a discrete set of values of the variable in the interval (discrete or digital sampling), or may include the continuum of its values throughout (0,T) (continuous or analog sampling). In either case it is convenient to consider the "value" of the composite variable ($\underset{\sim}{S}$, for example) as the aggregate of the values of the corresponding physical quantity (S) at the appropriate instants in time $t_1 \ldots t_n$, e.g., $\underset{\sim}{S} = (S_1, S_2, \ldots S_n)$ and represent it as a point in a space of corresponding dimensionality (n) in the usual fashion.

The occurrence of each value is assumed to be governed by a probability distribution function defined over the space. These may be described by density functions if the space is continuous, and if discrete, as the elementary probabilities associated with each value. Thus, if the signal space Ω contains two n-dimensional points only, $\underset{\sim}{S}_0$ and $\underset{\sim}{S}_1$ corresponding to known signals which occur with probabilities q and p, the distribution function $\sigma(\underset{\sim}{S})$ is defined as

$$\left.\begin{aligned} \sigma(\underset{\sim}{S}_0) &= q \\ \sigma(\underset{\sim}{S}_1) &= p \end{aligned}\right\} , \quad p + q = 1 . \tag{1.1}$$

If $\underset{\sim}{S}$ is a composite random variable, e.g., a noise wave where each of the components S_k of $\underset{\sim}{S}$ possesses a continuous distribution with possible correlation between components, then $\sigma(\underset{\sim}{S}) = \sigma(S_1, S_2, \ldots, S_n)$ is an n-fold joint distribution function and there are infinitely many points in the space Ω. Similar remarks hold for the variables $\underset{\sim}{N}, \underset{\sim}{V}, \underset{\sim}{\gamma}$ and their probability distributions $W(\underset{\sim}{N})$, $F_s(\underset{\sim}{N})$ and $\delta(\underset{\sim}{\gamma}|\underset{\sim}{V})$, respectively. The information about $\sigma(\underset{\sim}{S})$ and $W(\underset{\sim}{N})$ available in any particular problem may not be complete, of course; we may know merely that they belong to certain classes of distributions.

The decision $\underset{\sim}{\gamma}$ may be any statement concerning the member of the signal class Ω present at the input. In the binary* detection problem, for instance, we test the hypothesis H_0: $\underset{\sim}{S} = 0$ (noise

*We use here and elsewhere (1.21) the term "binary" or "single alternative" to include all cases of detection where only one signal on any one observation (0,T) can be present. This particular signal, however, may be drawn from one or more sub-ensembles, representing possible distributions over the parameter describing the signals, e.g., amplitude, durations, or other structure factors. Thus the "simple alternative" refers to an ensemble containing only one possible signal, while the term "one-sided alternative" refers to a signal selected from an ensemble where there is more than one possibility.

alone) against the alternative H_1: $\underline{S} \neq 0$ (signal plus noise), and the decision space Δ contains two points: γ_0: $\underline{S} = 0$ is present at the input, and γ_1: $\underline{S} = 0$ is present at the input. Here Ω may contain only one other point besides $\underline{S} = 0$ (simple alternative)*, or a continuum of points corresponding, say, to different signal amplitudes (one-sided alternative)*. In extraction, $\underline{\gamma}$ is an estimate of the value or numerical measure of $\underline{S}$ present at the input, so that the dimensionality of the two is the same.

The decision rule $\delta(\underline{\gamma} \mid \underline{V})$ is the (conditional) probability that $\underline{\gamma}$ will be decided when $\underline{V}$ is given. Ordinarily this probability is either one or zero for each $\underline{V}$ and $\underline{\gamma}$ (nonrandomized decision rule), although the possibility of using a random mechanism to obtain $\underline{\gamma}$ from $\underline{V}$ is not excluded in the general formulation. The essence of the decision situation is that $\delta(\underline{\gamma} \mid \underline{V})$ is to be a rule for making the decision $\underline{\gamma}$ from the received or a posteriori data $\underline{V}$ alone, i.e., independently of any a posteriori information about $\underline{S}$ [cf. (3.15)]* although, of course, a priori knowledge of $\underline{S}$ is necessarily "built into" the decision rule. We indicate this symbolically by

$$\delta(\underline{\gamma} \mid \underline{V}) = \delta(\underline{\gamma} \mid \underline{V}, \underline{S}) \quad . \tag{1.2}$$

The decision rule is the mathematical embodiment of the physical system to be optimized. Once the "best" decision rule is found for the problem at hand (consistent with an appropriate definition of "best"—which, in turn, means an appropriate choice of criterion—) the system is designed to perform the optimum operation thus revealed. Note that the decisions $\underline{\gamma}$ are terminal decisions, so that the decision rule may involve a sequence of intermediate "sub-decisions," as does, for example, the standard sequential test of an hypothesis.

The observation space Γ contains points corresponding to all sample values of the process $\underline{V}$. $F_S(\underline{V})$ is the probability distribution function of the $\underline{V}$'s for given $\underline{S}$. If the form of the noise distribution $W(\underline{N})$ is known, as we shall assume in this paper, $F_S(\underline{V})$ is uniquely specified by the value of $\underline{S}$, i.e., $F_S(V)$ is a parametric family of distribution functions. If the form of the noise distribution is unknown, however, $F_S(\underline{V})$ becomes a nonparametric family. The theory to follow in its most general form includes this case [1.22-1.24]. Moreover, to discuss specifically a wide class of practical problems we shall further assume here that the mixture of signal and noise is additive ($\underline{V} = \underline{S} + \underline{N}$) and that signal and noise are statistically independent.** Thus,

$$F_S(\underline{V}) = W(\underline{V} - \underline{S}) \tag{1.3}$$

The nature of the distributions $F_S(\underline{V})$ is of the utmost importance, of course, as it is upon these that specific calculations of performance depend. They are simplified if the individual sample values $V_1, V_2, \ldots, V_n$ (the "coordinates" of $\underline{V}$) are statistically independent, since then $F_S(\underline{V})$ factors into a product of identical one-dimensional distribution functions. When sampling is continuous and V(t) is band-limited (0, B) with $T \gg (2B)^{-1}$, the sampling theorem [1.19], [1.25] shows how V(t) may be expressed with good approximation in terms of a finite number of uncorrelated sample values. Observation space may then be taken as a space with these coordinates instead of the original function space of V(t). In many situations, however, the inherent correlations and the finiteness of the interval play a central rôle, and cannot be safely approximated away.

More generally, $F_S(\underline{V})$ and Γ may be replaced in Fig. 1 by a distribution $G_S(\underline{x})$ and a space X whose coordinates are functionals of the continuous process V(t), $0 \leq t \leq T$, [1.14, 1.27]. Thus, for example, if $\{\phi_\nu\}$ is a complete orthonormal set of functions satisfying

$$\phi_\nu(t) = \lambda_\nu \int_0^T R(s,t)\, \phi_\nu(s)\, ds \quad , \tag{1.4}$$

where R(s,t) is the correlation function of the process, then the coordinates defined by

$$x_\nu = \int_0^T V(t)\, \phi_\nu(t)\, dt \tag{1.5}$$

are uncorrelated random variables. When V(t) is a Gaussian process, x_ν is a Gaussian variable whose first two moments are easily found, so that $G_S(\underline{x})$ is immediately forthcoming. When V(t) is

*See footnote previous page

**While this simplifies the analysis in many cases, it in no way diminishes the scope of the general decision theory.

not Gaussian, however, the difficulties of finding distributions of functionals like x_ν become formidable [1.26]. Another and ultimately equivalent method of handling correlated samples is based on the expansion of the distribution $W(\underline{Y}-\underline{S})$ about $\underline{S} = 0$ [1.13]. This proves to be more useful for the threshold cases of interest here.* (See Sec. 3).

Figure 1 emphasizes that decision rules are essentially transformations that map observation space into decision space. In detection each point of the observation space Γ (or $\mathcal{X}$) is mapped into one or the other of the two points constituting the space of terminal decision Δ. In binary detection, this is the same as dividing observation space into two regions, one corresponding to "no signal" and the other to "signal plus noise" and carrying out the decision operation in one step, since only a single alternative is involved. The binary detection problem is then the problem of how best to make this division. Similarly, in extraction each point of observation space is mapped into a point of the space of terminal decisions Δ which in this instance has the same structure as the signal space Ω. If the dimensionality of Δ is smaller than that of Γ (as is usually the case in estimating a signal parameter) the transformation is "irreversible," i.e., many points of Γ go into a single point of Δ. Thus extraction may also be thought of as the division of Γ into regions. Detection and extraction are closely related, of course, since it is only necessary to group the points of Δ corresponding to $\underline{S} = 0$ into a single class labeled "signal and noise" to transform an extractor into a detector. Similarly, detection systems are often essentially extractors followed by a threshold device that separates $\underline{S} = 0$ from $\underline{S} \neq 0$. A system optimized for one function, however, may not necessarily be optimized for the other, and in this sense we may therefore consider detection and extraction as separate problems for analysis.

The scheme of Fig. 1 is clearly flexible enough to encompass a wide variety of reception problems. The problem is characterized by the information available about $\sigma(\underline{S})$ and $W(\underline{N})$, the mode of combination, $\oplus$, and the criterion chosen for the optimum decision rule. Here $F_{\underline{S}}(\underline{Y})$ (or $G_{\underline{S}}(\underline{x})$) and $\delta(\underline{\gamma} \mid \underline{Y})$ are derived quantities, the latter being the "answer" sought. In Section 2 we show how the risk and information-loss criteria lead to classes of optimum decision rules. Sections 3 and 4 apply these results to detection and extraction with some simple illustrative examples, while Section 5 discusses connections between the risk and information criteria. A short review of some of the new features of the present approach and its implications in statistical communication theory conclude the paper.

2. Summary of Main Statistical Methods

2.0. Introduction

The theory of statistical decision functions was founded by Abraham Wald [2.1]. Since its inception the theory has been the subject of intensive research by many mathematical statisticians. A recent book by Blackwell and Girshick [2.2] gives the present status with an up-to-date bibliography.

Our object here is to show how the concepts and results of this theory may be applied to practical communication problems, and may be made to furnish a reasonable basis for system design. The concept of the loss function used to date in decision theory is now generalized to include criteria of information loss, as well as risk, in a single formulation.

2.1. Evaluation Functions

The first requirement in a definition of optimum system performance is some kind of an evaluative scheme, or basis for saying that one system is better than another. We need a way of assigning an evaluation to each decision rule. In view of the statistical nature of the decision problem, this evaluation should depend on the long-run or "ensemble" performance of the system. Let $\mathcal{J}(S, \underline{\gamma})$ be a function which assigns a loss to each possible combination of signal and decision, and

*Although in the following discussion Γ, $\underline{Y}$ and $F_S(\underline{Y})$ will be used throughout, it should be understood that the results hold equally well with $\mathcal{X}$, X and $G_s(X)$ in their places, provided assumptions as to the continuity, etc. of $F_S(\underline{Y})$ are transferred to $G_s(\underline{X})$.

let $\mathcal{E}$ denote the operation of combining these to give the decision rule δ an over-all loss rating which takes account of all possible modes of behavior and their relative frequencies of occurrence. In this paper $\mathcal{E}$ will be taken as the expectation or average value of $\mathcal{F}$ ($\mathcal{E} \rightarrow E$), since this leads to useful interpretations of previous results as special cases of ours. The possibility of using other linear or nonlinear operations for $\mathcal{E}$ should not be overlooked, however.

The conditional loss rating of δ is then defined for given signals as:

$$r(\underset{\sim}{S},\delta) = E_F \mathcal{F}(\underset{\sim}{S},\underset{\sim}{Y}) = \int_{\Gamma}\int_{\Delta} \mathcal{F}(\underset{\sim}{S},\underset{\sim}{Y})\, F_s(\underset{\sim}{V})\ \delta(\underset{\sim}{Y}|\underset{\sim}{V})\, d\underset{\sim}{V} d\underset{\sim}{Y} \ . \tag{2.1}$$

The average loss rating of δ takes account of the a priori signal distribution:

$$\mathcal{V}(\sigma,\delta) = E_{F,\sigma}\mathcal{F}(\underset{\sim}{S},\underset{\sim}{Y}) = E_\sigma r(\underset{\sim}{S},\delta) = \int_{\Omega}\int_{\Gamma}\int_{\Delta} \mathcal{F}(\underset{\sim}{S},\underset{\sim}{Y})\, \sigma(\underset{\sim}{S}) F_s(\underset{\sim}{V})\ \delta(\underset{\sim}{Y}|\underset{\sim}{V})\, d\underset{\sim}{S}\ d\underset{\sim}{V} d\underset{\sim}{Y} \ . \tag{2.2}$$

The loss function $\mathcal{F}(\underset{\sim}{S},\underset{\sim}{Y})$ may or may not depend on how the decision is reached. In the following we shall consider one of each type, corresponding to the risk and information loss criteria. In risk theory a cost $C(\underset{\sim}{S},\underset{\sim}{Y})$ is preassigned to each combination of signal and decision independently of δ :

$$\mathcal{F} = C(\underset{\sim}{S},\underset{\sim}{Y}) \ . \tag{2.3}$$

The loss function for the information-loss criterion, however, is the "uncertainty" about $\underset{\sim}{S}$ when $\underset{\sim}{Y}$ is known, (or the "surprisal"), [p.g] defined as:

$$\mathcal{F} = -\log P_{(2)}(\underset{\sim}{S}|\underset{\sim}{Y}), \tag{2.4}$$

where $P_{(2)}(\underset{\sim}{S}|\underset{\sim}{Y})$ is the a posteriori probability of $\underset{\sim}{S}$ given $\underset{\sim}{Y}$. This clearly depends not only on $\underset{\sim}{S}$ and $\underset{\sim}{Y}$, but on the decision rule in use as well: it cannot be preassigned independently of δ .

The conditional and average loss ratings of δ follow from (2.1) and (2.2):

Conditional risk:	$r(\underset{\sim}{S},\delta) = E_F\, C(\underset{\sim}{S},\underset{\sim}{Y})$	(2.5)
Average risk:	$R(\sigma,\delta) = E_{F,\sigma}\, C(\underset{\sim}{S},\underset{\sim}{Y}) = E_\sigma\, r(\underset{\sim}{S},\delta)$	(2.6)
Conditional Information loss:	$h(\underset{\sim}{S},\delta) = -E_F \log P_{(2)}(\underset{\sim}{S}|\underset{\sim}{Y})$	(2.7)
Average Information loss:	$H(\sigma,\delta) = -E_{F,\sigma}\log P_{(2)}(\underset{\sim}{S}|\underset{\sim}{Y}) = E_\sigma\, h(\underset{\sim}{S},\delta)$.	(2.8)

The last of these is the well known "equivocation" in information theory [2.4].

2.2. Minimax and Bayes Criteria.

A central feature of the present theory is that the question of a priori probabilities*, avoided for the most part in classical statistical testing procedures and the subject of much controversy when considered [2.5], is here squarely faced, as indeed it must be for satisfactory solution of the practical problems of reception. These prior probabilities are often not known and often cannot be obtained (they may not even exist in a satisfactory philosophical sense). In that case a basis for design is provided by the Minimax criterion:

> A Minimax decision rule δ_o is one whose maximum conditional loss rating (over all signal values) is not greater than the maximum conditional loss rating of any other decision rule δ :

$$\operatorname*{Max}_{S} r(\underset{\sim}{S},\delta_o) \leq \operatorname*{Max}_{S} r(\underset{\sim}{S},\delta) \quad , \quad \text{all } \delta \ . \tag{2.9}$$

The Minimax principle may be criticized as being too conservative. When the decision situation is regarded as a game [2.6] between Nature and the statistician, for example, we observe that Minimax

*Here we distinguish between a priori probabilities $\sigma(\underset{\sim}{S})$, and a priori noise probabilities $W(\underset{\sim}{N})$ etc. Throughout the present paper however, we shall use the term a priori probability to refer to the former only; for $W(\underset{\sim}{N})$, see comment on nonparametric tests, Sec. (1.1).

strategy on the part of the latter is reasonable only if he assumes that Nature deliberately chooses for her strategy the one least favorable to him. Although this is not likely, there is perhaps some justification for basing a theory of the best choice of δ on this assumption, when nothing whatever is known about Nature's strategy. Even if the analogy with game theory is incomplete at this point, it nevertheless remains valuable, since it led Wald to the complete class theorem (Sec. 2.4).

When the prior (e.g., signal) probabilities ($\sigma(\underset{\sim}{S})$) are completely known, the Bayes criterion is the basis for design:

> A Bayes decision rule δ_σ is one whose average loss rating is smallest for a given a priori distribution $\sigma(S)$:

$$R(\sigma, \delta_\sigma) = \min_{\delta} R(\sigma, \delta) \quad . \tag{2.10}$$

The Bayes principle makes the fullest use of prior (signal) probabilities; in a sense it assumes the most favorable case. The two principles are appropriate in the extreme cases of no knowledge and full knowledge of these probabilities. For intermediate situations, where $\sigma(S)$ is partially known, see the discussion of Hodges and Lehmann [2.7]. We emphasize that when $W(\underset{\sim}{N})$ is not known the minimax principle can be used to supply a "least favorable" $W(\underset{\sim}{N})$ and corresponding Bayes test, just as it may be used to handle incomplete knowledge of a priori signal probabilities.

In either situation minimax procedures provide the necessary probability distribution with which to construct Bayes tests. See Sec. 3.4 following for an illustration.

2.3. Comparison of Decision Rules

The conditional loss rating of a decision rule depends, of course, on the particular signal present at the input. One decision rule may have a smaller rating than another for some signals, and a larger one for others. If the conditional loss rating of δ_1 never exceeds that of δ_2 for any value of $\underset{\sim}{S}$, and is actually less than that of δ_2 for some particular $\underset{\sim}{S}$, δ_1 is said to be uniformly better than δ_2. This leads to the definition of an admissible decision rule:

> A decision rule is admissible if no uniformly better one exists.

Note particularly that on this definition an admissible rule is not necessarily uniformly better than any other. Other rules can have smaller ratings at particular values of $\underset{\sim}{S}$. The point is that they cannot be better at all values of $\underset{\sim}{S}$.

It follows from these definitions that if a Bayes or Minimax rule is unique it is admissible. The converse is not true, however; an admissible rule is not necessarily Bayes or Minimax. Thus, no system that does not minimize the average loss rating can be uniformly better than a Bayes system (for the same σ), and no system that does not minimize the maximum conditional loss rating can be uniformly better than a Minimax system. Admissibility is an important additional optimum property of Bayes and Minimax decision systems.

It is clear from the definitions also that a Bayes decision rule whose conditional loss rating is constant is a Minimax decision rule. In many cases this furnishes a useful way of finding the Minimax rule. (See Sec. 3).

2.4. The Complete Class Theorem

A class D of decision rules is complete if for any δ not in D we can find a δ^* in D such that δ^* is uniformly better than δ. If D contains no sub-class which is complete it is a minimal complete class.

Wald has shown [2.8] that in the risk formulation [($\mathcal{F} = C(\underset{\sim}{S}, \underset{\sim}{Y})$], the class of all admissible Bayes decision rules (i.e., for different σ's) is a minimal complete class [2.9] under the following conditions [2.10]:

(A) $F_S(\underset{\sim}{Y})$ is continuous in $\underset{\sim}{S}$,
(B) $C(\underset{\sim}{S}, \underset{\sim}{Y})$ is bounded in $\underset{\sim}{S}$ and $\underset{\sim}{Y}$,
(C) The class of decision rules considered is restricted to either (i) nonsequential rules, or (ii) sequential rules, (2.11)
(D) $\underset{\sim}{S}$ and $\underset{\sim}{Y}$ are restricted to finite closed domains.

On these assumptions also, any Minimax decision rule can be shown to be a Bayes rule with respect to a certain $\sigma(S)$, called the least favorable a priori distribution, and the existence of such a distribution as well as the existence of Bayes and Minimax rules themselves is assured [2.11].

The complete class theorem thus establishes an optimum property of the Bayes class as a whole. For instance, we shall see in Sec. 3 that the Bayes test for binary detection is a likelihood ratio test. The complete class theorem applied here says that corresponding to any non-likelihood receiver (for example, most existing detection systems) there is a likelihood receiver which is uniformly better. The physical embodiment of such a receiver comprises a computer of the likelihood ratio and a threshold comparison device. The computer design is the same for all Bayes tests, so that the optimum property attributed to the Bayes class as a whole by the complete class theorem is realized in the computer design.

No complete class theorem for the information loss formulation has been proved as yet. Some results on the characterization of Bayes tests with this measure used in hypothesis testing and extraction are given in Sec. 5.

2.5. Information and Sufficiency.

Here we wish to make clear the connection between information loss and sufficient statistics. A statistic of the distribution $F_S(\underset{\sim}{Y}) = F(\underset{\sim}{Y}|\underset{\sim}{S})$ may be defined as any transformation of the observation points such as $\gamma(\underset{\sim}{Y})$. An estimator of $\underset{\sim}{S}$ is a statistic formed by mapping the points of Γ onto a decision space like Ω. The maximum likelihood estimator of $\underset{\sim}{S}$, for example, is constructed by giving to the estimator $\underset{\sim}{\gamma}$, for each $\underset{\sim}{Y}$, the value of $\underset{\sim}{S}$ that maximizes $F(\underset{\sim}{Y}|\underset{\sim}{S})$.

The classical concept of sufficiency is based on the view that the parameters $\underset{\sim}{S}$ governing the distribution exert a "causal" influence on the value of the variate $\underset{\sim}{Y}$ which is more or less obscured by the "randomness" of the distribution. A statistic $\gamma(\underset{\sim}{Y})$ is said to be sufficient, roughly speaking, if knowledge of $\underset{\sim}{\gamma}$ is as good as knowledge of $\underset{\sim}{Y}$ itself as far as determining the $\underset{\sim}{S}$ "responsible" for both is concerned. From (1.2), we write

$$\delta(\underset{\sim}{\gamma}|\underset{\sim}{Y}) = \delta(\underset{\sim}{\gamma}|\underset{\sim}{Y},\underset{\sim}{S}) \quad .$$

A little algebraic rearranging gives:

$$\sigma(\underset{\sim}{S})\, F(\underset{\sim}{Y}|\underset{\sim}{S}) = \sigma(\underset{\sim}{S}) P_3(\underset{\sim}{\gamma}|\underset{\sim}{S})\, \frac{\eta(\underset{\sim}{Y}|\underset{\sim}{S},\underset{\sim}{\gamma})}{\delta(\underset{\sim}{\gamma}|\underset{\sim}{Y})} \quad , \tag{2.12}$$

where $P_3(\underset{\sim}{\gamma}|\underset{\sim}{S})$ is the conditional probability (density) of $\underset{\sim}{\gamma}$ with $\underset{\sim}{S}$ fixed, and $\eta(\underset{\sim}{Y}|\underset{\sim}{S},\underset{\sim}{\gamma})$ that of $\underset{\sim}{Y}$ with $\underset{\sim}{S}$ and $\underset{\sim}{\gamma}$ fixed. The only kind of knowledge about $\underset{\sim}{S}$ we get by knowing $\underset{\sim}{Y}$ is contained in the dependence of $F(\underset{\sim}{Y}|\underset{\sim}{S})$ on $\underset{\sim}{S}$, $\underset{\sim}{Y}$ fixed, namely, the likelihood function. If we know $\underset{\sim}{\gamma}$, but not $\underset{\sim}{Y}$, we can reproduce this dependence (except for an unknown constant scale factor) from a knowledge of $\underset{\sim}{\gamma}$ alone, only in case $\eta(\underset{\sim}{Y}|\underset{\sim}{S},\underset{\sim}{\gamma})$ is independent of $\underset{\sim}{S}$, i.e., if

$$\eta(\underset{\sim}{Y}|\underset{\sim}{S},\underset{\sim}{\gamma}) = \eta(\underset{\sim}{Y}|\underset{\sim}{\gamma}) \quad . \tag{2.13}$$

With this, (2.12) becomes equivalent to

$$P_1(\underset{\sim}{S}|\underset{\sim}{Y}) = P_2(\underset{\sim}{S}|\underset{\sim}{\gamma}) \quad , \tag{2.14}$$

where $P_1(\underset{\sim}{S}|\underset{\sim}{Y})$ and $P_2(\underset{\sim}{S}|\underset{\sim}{\gamma})$ are the conditional probability densities of $\underset{\sim}{S}$ with $\underset{\sim}{Y}$ and $\underset{\sim}{\gamma}$ respectively fixed. Either (2.13) or (2.14) may be taken as the definition of sufficiency. When $\underset{\sim}{\gamma}$ is a sufficient statistic, specification of $\underset{\sim}{Y}$ in addition to $\underset{\sim}{\gamma}$ does not in any way improve our knowledge of $\underset{\sim}{S}$. The relation (2.12) shows that only distributions that can be factored into a product of two terms, such that one involves $\underset{\sim}{\gamma}$ and $\underset{\sim}{S}$ only and the other $\underset{\sim}{Y}$ and $\underset{\sim}{\gamma}$ only, admit a sufficient statistic. The notion of sufficiency and the factorization condition were introduced by R.A. Fisher [2.12]. A recent treatment has been given by Halmos and Savage [2.13].

Closely related to sufficiency is the idea, due also to Fisher [2.14], of associating with an observation a numerical measure of the information it contains about the distribution parameter. The Shannon information measure serves the same purpose and is more appropriate for communication problems. We define the loss of information about a particular $\underset{\sim}{S}$ attending formation of the statistic or estimator $\underset{\sim}{\gamma}$, from a particular observation $\underset{\sim}{Y}$, as the difference of the uncertainties, or

$$\log \frac{P_1(\underset{\sim}{S}|\underset{\sim}{Y})}{P_2(\underset{\sim}{S}|\underset{\sim}{\gamma})} \quad , \tag{2.15}$$

The expected value of this, or the average information loss, is

$$H(\sigma,\delta) = \int_{\Delta}\int_{\Gamma} f(\underline{V})\, d\underline{V}\, d\underline{\gamma}\ \delta(\underline{\gamma}|\underline{V}) \int_{\Omega} P_1(\underline{S}|\underline{V}) \log \frac{P_1(\underline{S}|\underline{V})}{P_2(\underline{S}|\underline{\gamma})}\, d\underline{S}\ , \tag{2.16}$$

where

$$f(\underline{V}) = \int_{\Omega} \sigma(\underline{S})\, F_s(\underline{V})\, d\underline{S}\ . \tag{2.17}$$

The last integral in (2.16) is always positive or zero, and is zero if and only if (2.14) is satisfied, i.e., if $\underline{\gamma}$ is a sufficient statistic [2.15 , 2.16].

3. Application to Detection

3.0. Introduction

In this section we consider detection mainly from the risk point of view, and examine explicitly the binary, or single-alternative cases. Some results of the information-theory approach are given in Section 5.

The binary detection problem [1.21], [1.25] is the problem of testing the hypothesis H_0: noise alone present at the input–against the alternative H_1: signal plus noise present,–when there are only two points in decision space, namely, γ_0: the decision that noise alone occurs, and γ_1: the decision that signal plus noise occurs; γ_0 and γ_1 are decided with probabilities $\delta(\gamma_0|\underline{V})$ and $\delta(\gamma_1|\underline{V})$ respectively, and

$$\delta(\gamma_0|\underline{V}) + \delta(\gamma_1|\underline{V}) = 1\ . \tag{3.1}$$

A cost is preassigned to each possible combination of signal and decision as follows:

$$\begin{aligned} C(\underline{S}=0,\ \gamma_0) &= C_{1-\alpha} & \qquad C(\underline{S}=0,\ \gamma_1) &= C_{\alpha} \\ C(\underline{S}\neq 0,\ \gamma_0) &= C_{\beta} & \qquad C(\underline{S}\neq 0,\ \gamma_1) &= C_{1-\beta}\ , \end{aligned} \tag{3.2}$$

where C_{α} and C_{β} are the costs of the Type I and Type II errors mentioned previously, and $C_{1-\beta}$ and $C_{1-\alpha}$ are the costs of correct decisions. These are all finite positive quantities with

$$C_{\alpha} > C_{1-\alpha} \quad , \quad C_{\beta} > C_{1-\beta} \tag{3.3}$$

as required by the nature of the problem. We also specify that $W(\underline{V}-\underline{S})$ is continuous in $\underline{S}$, and consider nonsequential tests only, thus fulfilling all of the assumptions of Sec. 2.4.

We next consider the signal space Ω to contain infinitely many points besides $\underline{S} = 0$, with $\sigma(S)$ taken for convenience as [1.21]

$$\begin{aligned} &\sigma(S) = q\,\delta(\underline{S}-0) + w(\underline{S}) \\ &w(0) = 0 \qquad ;\ p = \int_{\Omega} w(\underline{S})\, d\underline{S} \\ &p + q = 1\ . \end{aligned} \tag{3.4}$$

Thus the test is against a one-sided alternative, the simple alternative test appearing as the special case

$$w(\underline{S}) = p\,\delta(\underline{S}-\underline{S}_1)\ , \tag{3.5}$$

where S_1 is the single possible signal.

Note that no specific assumption is made as to the way different signals are distinguished in Ω.

They may all be of the same form with different amplitudes or epochs or both, or they may differ in other respects. This detail is reflected in the expression used for $w(\underline{S})$, which is a one-dimensional distribution if amplitude alone varies, two-dimensional if both amplitude and epoch vary, n-dimensional if the values of n physical quantities are used to specify the signal completely in the interval (0,T), etc. The following discussion includes all of these.

3.1. Characterization of Bayes Detection Systems.

To characterize the class of Bayes tests, we form the expression for the average risk and minimize by a choice of $\delta(\gamma_o|\underline{V})$ or $\delta(\gamma_1|\underline{V})$. The conditional risk depends on whether $\underline{S} = 0$ or $\underline{S} \neq 0$ is present at the input. From (2.5) we write

$$r(S,\delta) = \int_{\Gamma}\int_{\Delta} C(\underline{S},\underline{\gamma})\, W(\underline{V}-\underline{S})\, \delta(\underline{\gamma}|\underline{V})\, d\underline{V}\, d\underline{\gamma}$$

$$= \left.\begin{array}{ll} \alpha\, C_{\alpha} + (1-\alpha)\, C_{1-\alpha} & , \quad \underline{S} = 0 \\ \beta\, C_{\beta} + (1-\beta)\, C_{1-\beta} & , \quad \underline{S} \neq 0 \end{array}\right\} \tag{3.6}$$

where α and β are the (conditional) probabilities of Type I and Type II errors respectively:

$$\left.\begin{array}{l} \alpha \equiv \int_{\Gamma} W(\underline{V})\, \delta(\gamma_1|\underline{V})\, d\underline{V} \quad , \\ \beta \equiv \int_{\Gamma} W(\underline{V}-\underline{S})\, \delta(\gamma_o|\underline{V})\, d\underline{V} \quad . \end{array}\right\} \tag{3.7}$$

The average risk takes account of $\sigma(\underline{S})$, given by (3.4). From (2.6) and (3.1) we then have

$$R(\sigma,\delta) = q\alpha C_{\alpha} + q(1-\alpha) C_{1-\alpha} + \beta' C_{\beta} + (p-\beta') C_{1-\beta} \tag{3.8}$$

$$= \int_{\Gamma} \left\{ q C_{\alpha} W(\underline{V}) + C_{1-\beta} \langle W(\underline{V}-\underline{S})\rangle + \delta(\gamma_o|\underline{V}) \left[\langle W(\underline{V}-\underline{S})\rangle (C_{\beta} - C_{1-\beta}) - qW(\underline{V})(C_{\alpha} - C_{1-\alpha}) \right] \right\} d\underline{V}, \tag{3.9}$$

where now

$$\langle W(V-S)\rangle \equiv \int_{\Omega} W(\underline{V}-\underline{S})\, w(\underline{S})\, d\underline{S} \quad , \tag{3.10}$$

$$\beta' \equiv \int_{\Gamma} \langle W(\underline{V}-\underline{S})\rangle\, \delta(\gamma_o|\underline{V})\, d\underline{V}. \tag{3.11}$$

Note that in the simple alternative case (3.5)

$$\left.\begin{array}{l} \langle W(\underline{V}-\underline{S})\rangle \longrightarrow pW(\underline{V}-\underline{S}_1) \quad , \\ \beta' \longrightarrow p\beta \quad . \end{array}\right\} \tag{3.12}$$

The condition for minimum $R(\sigma,\delta) = R(\sigma,\delta_{\sigma})$ is now evident from (3.9). Let:

$$\Lambda \equiv \frac{\langle W(\underline{V}-\underline{S})\rangle}{qW(\underline{V})} \quad , \tag{3.13}$$

$$K \equiv \frac{C_{\alpha} - C_{1-\alpha}}{C_{\beta} - C_{1-\beta}} \, . \tag{3.14}$$

Then the Bayes decision rule is:

$$\left.\begin{array}{l} \text{Let } \delta(\gamma_o|\underline{V}) = 0 \text{ (i.e., decide } \gamma_1) \text{ when } \Lambda > K \\ \text{Let } \delta(\gamma_o|\underline{V}) = 1 \text{ (i.e., decide } \gamma_o) \text{ when } \Lambda < K \end{array}\right\} . \tag{3.15}$$

The Bayes decision rule therefore divides the observation space Γ into two regions separated by the

$\underset{\sim}{V}$'s satisfying Λ = the threshold K, viz., the acceptance region Γ' containing $\underset{\sim}{V}$'s for which $\Lambda < K$, and the critical region Γ'' in which $\Lambda > K$. Note that the division is different for different σ's. The Bayes class of decision rules includes all rules of the type (3.15) corresponding to different a priori signal and noise probabilities.

Since (3.13) is a generalization of the likelihood ratio* this means that Bayes decision rules are nonrandom likelihood ratio tests with a threshold K depending on the preassigned costs. If a Bayes test is unique, it is admissible, which means here (according to (3.6)) that both of the error probabilities α and β cannot be smaller for any nonlikelihood test than they are for a likelihood ratio test. The complete class theorem then says that given any nonlikelihood ratio test one can always find a likelihood ratio test for which at least one of the error probabilities is less than, and the other is not greater than, those of the nonlikelihood test.

The likelihood ratio test takes on added significance when the Bayes risk (3.9) is rewritten as

$$R(\sigma,\delta) = \int_{\Gamma} f(\underset{\sim}{V}) \left[\rho_{\gamma_1}(\underset{\sim}{V})\, \delta(\gamma_1 | \underset{\sim}{V}) + \rho_{\gamma_0}(\underset{\sim}{V})\, \delta(\gamma_0 | \underset{\sim}{V}) \right] d\underset{\sim}{V} \quad , \tag{3.16}$$

where

$$f(V) = \int_{\Omega} \sigma(\underset{\sim}{S})\, W(\underset{\sim}{V}-\underset{\sim}{S})\, d\underset{\sim}{S} = q\, W(\underset{\sim}{V}) + \langle W(\underset{\sim}{V}-\underset{\sim}{S}) \rangle \tag{3.17a}$$

$$\rho_{\gamma 1}(\underset{\sim}{V}) = C_{\alpha} P_r(\underset{\sim}{S}=0 | \underset{\sim}{V}) + C_{1-\beta} P_r(\underset{\sim}{S}\neq 0 | \underset{\sim}{V}) \tag{3.17b}$$

$$\rho_{\gamma 0}(\underset{\sim}{V}) = C_{\beta} P_r(\underset{\sim}{S}\neq 0 | \underset{\sim}{V}) + C_{1-\alpha} P_r(\underset{\sim}{S}=0 | \underset{\sim}{V}) \quad , \tag{3.17c}$$

and where we have used the relations

$$q\, W(\underset{\sim}{V}) = f(\underset{\sim}{V})\, P_r(\underset{\sim}{S}=0 | \underset{\sim}{V}), \tag{3.18a}$$

$$\langle W(\underset{\sim}{V}-\underset{\sim}{S}) \rangle = f(\underset{\sim}{V})\, P_r(\underset{\sim}{S}\neq 0 | \underset{\sim}{V}). \tag{3.18b}$$

Here $f(\underset{\sim}{V})$ is the total probability of $\underset{\sim}{V}$'s occurrence, and $P_r(\underset{\sim}{S}=0 | \underset{\sim}{V})$ the posterior probability of $\underset{\sim}{S} = 0$ when $\underset{\sim}{V}$ is given, etc. Thus, for example, (3.18a) gives alternative expressions for the joint probability of $\underset{\sim}{S} = 0$ and $\underset{\sim}{V}$. The quantities $\rho_{\gamma 0}(\underset{\sim}{V})$ and $\rho_{\gamma 1}(\underset{\sim}{V})$ may be interpreted as the a posteriori risks of making decisions γ_0 and γ_1, respectively. The decision rule (3.15) which minimizes the Bayes risk thus also makes the decision for which the a posteriori risk is least.

Similarly, if we note that the logarithm of the likelihood ratio is proportional to the Shannon measure of the difference between the uncertainties about H_0 and H_1 when $\underset{\sim}{V}$ is known:

$$\log \Lambda = \log \frac{\langle W(\underset{\sim}{V}-\underset{\sim}{S}) \rangle}{q\, W(\underset{\sim}{V})} = \log \frac{P_r(\underset{\sim}{S}\neq 0 | \underset{\sim}{V})}{P_r(\underset{\sim}{S}=0 | \underset{\sim}{V})} \tag{3.19}$$

we observe that the decision rule (3.15) amounts to deciding in favor of H_1 when the uncertainty about H_1 is less than the uncertainty about H_0 by an amount log K. When there is only one signal (simple alternative case), which is distinguished from zero signal by an amplitude scale factor a_0, it is easy to show that the average of the uncertainty difference (i.e., the information difference) is proportional [3.1] to a_0^2, when a_0 is small.

3.2. The Neyman-Pearson and Ideal Observers [1.13], [1.18], [1.21]

The classical Neyman-Pearson test of an hypothesis against a single alternative fixes the probability α and minimizes the probability β. This is accomplished by a likelihood ratio test with a certain threshold which depends on α. To show this, one may use expressions (3.7), minimizing β by variation of the boundary between the acceptance and critical regions, subject to a constraint of fixed

*Here and elsewhere we use "likelihood ratio" and Λ to refer to the ratio of the joint probability of $\underset{\sim}{S} \neq 0$ and $\underset{\sim}{V}$ to that of $\underset{\sim}{S} = 0$ and $\underset{\sim}{V}$. This differs from the classical use of the term (as the ratio of the corresponding conditional probabilities) in connection with tests not concerned with the prior probabilities of the hypotheses. Note that with our definition Λ is essentially a ratio of a posteriori probabilities. (See Eq. 3.19).

α[1.18]. Since the Neyman-Pearson test is a likelihood ratio test, it may be interpreted from the risk point of view as a Bayes test for certain cost assumptions. Fixing α and minimizing β is thus equivalent to assuming a certain ratio of costs (depending on α, i.e., $K = K(\alpha)$), and minimizing the average risk. The smaller the allowed α (false alarm probability) the higher the threshold required; or in risk terms: the smaller α is the larger must be the ratio of costs preassigned to false alarms and false rests.

Another way of designing a single alternative test is to require that the total probability of error ($q\alpha+p\beta$) be minimized. An observer who makes decisions in this way is called an Ideal Observer. In a way similar to that used for the Neyman Pearson test, this also may be set up as a variational problem (with no constraint) resulting in a likelihood ratio test with $K = 1$ [1.18]. Thus it may be thought of as a Bayes test with a cost ratio of unity.

The fact that these are likelihood ratio tests follows from the optimum performance they require; since they are likelihood ratio tests, they belong to the Bayes class from the risk point of view, and therefore share the general optimum properties possessed by that class.

3.3. Decision Curves

The Bayes risk itself may be taken as a figure of merit for optimum system performance. In the simple alternative case, when it is known that there is only one other signal besides $\underset{\sim}{S} = 0$ in Ω, and it is characterized by a fixed amplitude scale factor a_o, one can define the minimum detectable signal (amplitude) in a way analogous to the betting curve procedure introduced by Siegert. The Bayes risk for detection depends on the signal amplitude, naturally being less for larger amplitudes. Thus, a functional relation between these quantities may be used to find the smallest signal amplitude for which the risk does not exceed a certain value (assumed beforehand as a criterion for the minimum detectable signal). Here α and β may be calculated from the formulas:[1.18, 1.21]

$$\alpha = 1 - \int_{-\infty}^{\log K} dy \int_{\Gamma} W(\underset{\sim}{V})\ \delta(y - \log \Lambda(\underset{\sim}{V}))\ d\underset{\sim}{V}\ , \tag{3.20}$$

$$\beta = 1 - \int_{\log K}^{\infty} dy \int_{\Gamma} W(\underset{\sim}{V} - a_o \underset{\sim}{S})\ \delta(y - \log \Lambda(\underset{\sim}{V}))\ d\underset{\sim}{V} \tag{3.21}$$

where $\underset{\sim}{S}$ is now normalized to place the amplitude factor a_o in evidence [1.13], and δ is the Dirac delta function. Here $\Lambda(\underset{\sim}{V})$, or any monotonic function of $\Lambda(\underset{\sim}{V})$, may be used in place of $\log \Lambda(\underset{\sim}{V})$ for convenience and K is simply a threshold value.

Figure 2 shows normalized Bayes risk curves of this nature, calculated for Rayleigh statistics. Observation space in this example is the space of all values of the envelope of the mixture of signal and noise, or noise alone (See Sec. 1.1). Thus, $W(\underset{\sim}{V})$ and $W(\underset{\sim}{V} - a_o \underset{\sim}{S})$ in (3.20) and (3.21) are replaced by the corresponding Rayleigh distribution functions for noise alone and signal plus noise. It is assumed that n observations of the envelope are made in (0,T) at a repetition period large compared with the correlation time of the input, so that the samples are uncorrelated [3.2]. For the common cases of small a_o and large n, (3.20) and (3.21) yield approximately:

$$\alpha = \frac{1}{2}\left\{1 - \Phi\left[\frac{a_o^2\sqrt{n}}{2\sqrt{2}} + \frac{\log(K/\mu)}{a_o^2\sqrt{2n}}\right]\right\}, \tag{3.22}$$

$$= \frac{1}{2}\left\{1 - \Phi\left[\frac{a_o^2\sqrt{n}}{2\sqrt{2}} - \frac{\log(K/\mu)}{a_o^2\sqrt{2n}}\right]\right\} \tag{3.23}$$

where $\mu \equiv p/q$. Here

$$\Phi(z) \equiv (2/\sqrt{\pi})\int_0^z e^{-t^2}\, dt, \text{ as usual.}$$

Figure 3 summarizes these relations for $\alpha + \beta < 1$. For $a_o^2\sqrt{2n}$ fixed, the values of α and β are interchanged by reciprocating Λ/μ. Both α and β are decreased when $a_o^2\sqrt{2n}$ is increased with Λ/μ fixed. Specification of any two of the four quantities α, β, $a_o^2\sqrt{2n}$ and Λ/μ fixes the other two.

Thus, for example, an α of 0.15 and a β of 0.20 may be obtained only with $\mathcal{L}/\mu = 1.2$ and $a_o^2\sqrt{2n} = 2.7$.

The normalization of the Bayes risk curves of Fig. 2 is accomplished by noting that

$$R(\mu, \delta_\mu) \to q C_{1-\alpha} + p C_{1-\beta} \quad \text{as} \quad a_o \to \infty \ , \tag{3.24}$$

and as $a_o \to 0$,

$$R(\mu, \delta_\mu) \longrightarrow \begin{cases} q C_{1-\alpha} + p C_\beta \ , & \text{when} \quad K/\mu > 1 \\ q C_\alpha + p C_{1-\beta} \ , & \text{when} \quad K/\mu < 1 \ . \end{cases} \tag{3.25}$$

Thus, with the help of (3.24), (3.25), and (3.8), the normalized Bayes risk curves are here defined by

$$\mathcal{R}(\mu, \delta_\mu) = \frac{R(\mu, \delta_\mu) - (q C_{1-\alpha} + p C_{1-\beta})}{(q C_{1-\alpha} + p C_\beta) - (q C_{1-\alpha} + p C_{1-\beta})} = \alpha \frac{K}{\mu} + \beta \ , \quad (\frac{K}{\mu} > 1) \tag{3.26a}$$

$$\mathcal{R}(\mu, \delta_\mu) = \frac{R(\mu, \delta_\mu) - (q C_{1-\alpha} + p C_{1-\beta})}{(q C_\alpha + p C_{1-\beta}) - (q C_{1-\alpha} + p C_{1-\beta})} = \alpha + \frac{\mu}{K}\beta \ , \quad (\frac{K}{\mu} < 1) \tag{3.26b}$$

$$\mathcal{R}(\mu, \delta_\mu) = \alpha + \beta \ , \quad (\frac{K}{\mu} = 1) \ . \tag{3.26c}$$

Due to the symmetry of the pairs (3.22), (3.23) and (3.26a), (3.26b), the normalized risk for a given $a_o^2\sqrt{n}$ is the same for K/μ and μ/K when these have the same value.

The curves of Fig. 2 show that the minimum detectable signal, corresponding to a fixed fraction of maximum risk, is smallest when the cost ratio K is equal to the ratio μ of prior probabilities. Thus when μ is one, the Ideal Observer, who takes K = 1, minimizes the risk. The Neyman-Pearson tests ($K \neq 1$) for fixed μ and fixed risk yield smaller minimum detectable signals as the cost ratio is increased when $K < \mu$, and larger ones as it is increased when $K > \mu$.

Note from Fig. 2 that when $\mu \neq 1$ certain Neyman-Pearson tests can result in a smaller minimum detectable signal than the Ideal, if normalized risk is taken as the criterion, and when it makes sense to compare the two types of observer. As an example, suppose $\mu = 1/4$ and $a_o^2\sqrt{2n} = 2$. An Ideal Observer always takes $\mathcal{L} = K = 1$, which in this case (from Figures 2 and 3) yields a normalized Bayes risk of 0.79 with $\alpha = .045$ and $\beta = 0.61$. On the other hand, a Neyman-Pearson Observer who holds the false alarm probability α at 0.1 and minimizes β, needs $K/\mu = 2.3$ and obtains a minimized β of 0.45 (from Fig. 3). The latter's normalized Bayes risk is 0.68 from Fig. 2, and thus is smaller than the Ideal Observer's. However, (assume $C_{1-\alpha} = C_{1-\beta} = 0$; $K = C_\alpha / C_\beta$) the Ideal Observer with K = 1 weights Type I and Type II errors equally, while the Neyman Pearson Observer with K = 2.3/4 weights Type II errors more than Type I errors. The unnormalized risks here, given by $R(\mu, \delta_\mu) = p C_\beta \mathcal{R}(\mu, \delta_\mu)$, are $0.16\, C_{\beta_I}$ for the Ideal, and $0.14\, C_{\beta_{NP}}$ for the Neyman Pearson.

The comparison thus depends on the relative costs assigned to the two types of errors, by the two observers, i.e., if $C_{\beta_I} = C_{\beta_{NP}}$ (so that $C_{\alpha_I} = (4/2.3) C_{\alpha_{NP}}$), the Neyman Pearson risk is less than the Ideal, but if $C_{\alpha_I} = C_{\alpha_{NP}}$ (so that $C_{\beta_I} = (2.3/4) C_{\beta_{NP}}$) then the Ideal risk is less than the Neyman Pearson. The two observers may be compared on some other basis than risk, of course, such as the probability of a correct decision ($= 1 - q\alpha - p\beta$)[3,3]. For the example above, this is 0.842 for the Ideal and 0.830 for the Neyman Pearson, so that the Ideal is better in this respect, as it must be by definition, when percentage of correct decisions is chosen as the measure.

3.4. The Minimax Detection Rule

The theory of Sec. 2.4 shows that the Minimax rule provides the likelihood ratio test corresponding to a certain least favorable distribution $\sigma(S)$. To find the latter we take advantage of the fact that a likelihood ratio test with the same conditional risk for both hypotheses is Minimax.

As an example, let us now apply the procedure in the simple alternative case, where p and q (= 1-p) are unknown. We begin by varying p and q, calculating α and β for each likelihood ratio test thus defined, until an α and β are found for which the two risks of (3.6) are the same, i.e.,

$$\alpha C_\alpha + (1-\alpha) C_{1-\alpha} = \beta C_\beta + (1-\beta) C_{1-\beta} \ . \tag{3.27}$$

The minimax p and q (i.e., p_{mx}, q_{mx}) for which the right combination of α and β (i.e., α_{mx}, β_{mx}) occurs is then the least favorable a priori distribution.

No other test can give a smaller maximum conditional risk than the Minimax. Since the average risk cannot exceed the maximum conditional risk, the Minimax rule also has a smaller maximum average risk than any other (as p and q are varied). Of course, for some particular $p = p'$ and $q = q'$ another test might have smaller risk than the Minimax, for the same p' and q'. On the other hand, there would be a p and q for which it has a larger average risk. The Minimax rule thus has the advantage that it guards against the worst case, and the disadvantage that in doing so it admits the possibility of being bettered in particular cases.

Figure 4 shows some Minimax calculations for the example discussed above, with $C_{1-\alpha} = C_{1-\beta} = 0$ and $C_{\alpha} = 100$.

Because of these cost assumptions the curve labeled Minimax coincides with the Bayes risk curve for $\mu_{mx} = \mu = 1$ (i.e., $\mu = 1$ is the least favorable distribution here). The curves for other values of μ are Bayes risk curves, where $\sigma(S)$ is known a priori.

When $a_o^2 \sqrt{n}$ is 1 and $\mu = 8$, for example, we note that knowledge of μ reduces the risk by 20 units. Alternatively, if the minimum detectable signal is here defined as the smallest one for which the risk does not exceed 10, we find that $(a_o^2 \sqrt{n})_{min}$ is 1.40 when μ is known and 2.55 when μ is unknown. For fixed sample size n this means that the minimum detectable signal is increased by $10 \log(2.55/1.40) = 2.61$ db by lack of knowledge of μ. If the amplitude a_o is fixed, on the other hand, the integration time is increased by $\sqrt{2.55/1.40} = 1.35$, or by 35 per cent.

3.5. System Comparison

A likelihood-ratio receiver is in general a rather complex computer [1.13, 1.25]. It is important, therefore, to be able to compare the performance of a compromise, nonoptimum system with the optimum, so that the cost of the compromise may be a factor in system planning. Following ref. [1.21], Sec. 2, we observe in binary detection that just as the optimum system computes the likelihood ratio and compares the result with a certain threshold K, we may assume that the nonoptimum system also computes some quantity $F(\underset{\sim}{V};\underset{\sim}{S})$ for comparison with a threshold K, deciding "signal plus noise" when $F(\underset{\sim}{V};\underset{\sim}{S}) > K$ and "noise alone" when $F(\underset{\sim}{V};\underset{\sim}{S}) < K$. If the system function $F(\underset{\sim}{V};\underset{\sim}{S})$ is known, the probabilities α'_o and β'_o may be calculated from (3.20) and (3.21) with $F(\underset{\sim}{V};\underset{\sim}{S})$ in place of Λ and with the same threshold K.

As an example, we consider the coherent detection of a signal (of fixed amplitude) in Gaussian noise with continuous sampling in (0,T). Middleton [1.13] shows that the likelihood ratio becomes for this:

$$\log \Lambda = a_o \overline{\Phi(v,s)} - \frac{1}{2} a_o^2 \overline{\Phi(s,s)} \quad , \tag{3.28}$$

with

$$\overline{\Phi(v,s)} = \psi \int_0^T v(t)\, X(t)\, dt \quad , \tag{3.29a}$$

$$s(t) = \int_0^T X(s)\, K(t-s)\, ds \quad , \qquad (0- < t < T+). \tag{3.29b}$$

Here s and v are the normalized quantities: $S = a_o / \sqrt{\psi} s$; $V = \sqrt{\psi}\, v$, where ψ is the mean square noise amplitude, and K(t-s) is the auto-correlation function of the noise. When (3.28) is used in (3.20) and (3.21), the resulting α and β are exactly

$$\alpha = \frac{1}{2} \left\{ 1 - \Phi \left[\frac{a_o \sigma}{2} + \frac{\log(K/\mu)}{2 a_o \sigma} \right] \right\} \quad , \tag{3.30a}$$

$$\beta_{NP} = \frac{1}{2} \left\{ 1 - \Phi \left[\frac{a_o \sigma}{2} - \frac{\log(K/\mu)}{2 a_o \sigma} \right] \right\} \quad , \tag{3.30b}$$

where

$$\sigma^2 = \overline{\Phi(s,s)}. \tag{3.30c}$$

For the specific case of a sine wave of angular frequency ω_o, coherently detected with epoch $\epsilon_o = 0$, in broad-band Gaussian noise, shaped by an RC filter, $[(RC)^{-1} = \omega_F]$, we find that for a nonideal system which treats the noise as having an infinitely wide spectrum, i.e., no coherence from sample point to sample point, the minimum detectable signal, $(a_o^2)_{min}$, is 4.0 db higher than that obtained by the limiting or ideal system when the correlation in the noise is taken into account; $[(a_o^2)_{min}$ is chosen as a_o^2 for the 90% level of successful decisions; p = q = 1/2.] See Fig. 8 and (*iii*), Sec. 4 of reference

[1.13]. Here $F(\underset{\sim}{V};\underset{\sim}{S})$ replaces Λ in (3.28), or equivalently, $\overline{\Phi(s,s)}$, Eq. (3.30c) is replaced by the corresponding expression $\overline{\Phi(s,s')}$ in the nonideal situation.

4. Application to Extraction

4.0. Introduction

In this section we consider extraction as the counterpart of parameter estimation. Point estimation rather than estimation by confidence intervals is treated. Our main object here is to show how some of the methods commonly used for optimum extraction appear from the risk point of view [4.1].

As before, we let $\underset{\sim}{\gamma}$ denote the decision to be made about the signal $\underset{\sim}{S}$, and observe that when $\underset{\sim}{\gamma}$ is to be an estimate of $\underset{\sim}{S}$, the spaces Ω and Δ of Figure 1 have the same structure. We also assume that each contains a continuum of points and is a finite closed region which may, however, be taken large enough to be essentially infinite for practical purposes.

The cost function $C(\underset{\sim}{S},\underset{\sim}{\gamma})$ to be used in the risk analysis is, of course, to be preassigned in accordance with the external constraints of the problem, and is critical in determining the nature of the resulting system. A theorem of Hodges and Lehmann [4.2] says that

> if Δ is the real line and $C(\underset{\sim}{S},\underset{\sim}{\gamma})$ is a convex function* of $\underset{\sim}{\gamma}$ for every $\underset{\sim}{S}$, then for any decision rule δ there exists a nonrandomized decision rule whose risk is not greater than that of δ for all $\underset{\sim}{S}$ in Ω.

The squared-error cost function $C(\underset{\sim}{S},\underset{\sim}{\gamma}) = (\underset{\sim}{S}-\underset{\sim}{\gamma})^2$ is suitable for our purposes here, since it leads to conventionally used extraction procedures, and is also convex, so that the inconvenience of considering both randomized and nonrandomized rules may be avoided, at least in the one-dimensional case.**

A further simplification results from the fact that a nonrandomized decision (e.g., one for which is either 1 or 0) rule, may be written as:

$$\delta(\underset{\sim}{\gamma}|\underset{\sim}{V}) = \delta(\underset{\sim}{\gamma}-\gamma_s(\underset{\sim}{V})) \tag{4.1}$$

where the δ on the right is now the Dirac delta function. Here it is essential to distinguish between the estimate $\underset{\sim}{\gamma}$ and the estimator $\gamma_s(\underset{\sim}{V})$. The latter denotes the functional operation performed on the data $\underset{\sim}{V}$ by the system, while the former is simply a value of the output. The operation $\gamma_s(\underset{\sim}{V})$ is thus the quantity to be optimized.

4.1. Characterization of Bayes Extraction

To find the Bayes estimator, we begin with the conditional risk. From (2.5) and

$$r(\underset{\sim}{S},\gamma_s) = \int_{\Gamma}\int_{\Delta} C(\underset{\sim}{S},\underset{\sim}{\gamma})\, F_s(\underset{\sim}{V})\, \delta(\underset{\sim}{\gamma}-\gamma_s(\underset{\sim}{V}))\, d\underset{\sim}{V}\, d\underset{\sim}{\gamma} \tag{4.2}$$

$$= \int_{\Gamma} C[\underset{\sim}{S},\gamma_s(\underset{\sim}{V})]\, F_s(\underset{\sim}{V})\, d\underset{\sim}{V} \quad . \tag{4.3}$$

A useful alternative form, analogous to (3.6), is obtained by rearranging (4.2):

*A real-valued function $\psi(x)$ is convex in an interval (a,b) if for any x and y in (a,b), and any number $0 < \lambda < 1$, $\lambda\psi(x) + (1-\lambda)\psi(y) \geq \psi[\lambda x + (1-\lambda)y]$.

**When $\underset{\sim}{S}$ and $\underset{\sim}{\gamma}$ are multidimensional vectors, $(\underset{\sim}{S}-\underset{\sim}{\gamma})^2$ is to be interpreted as the length of the difference vector. Although the theorem above applies only to one-dimensional $\underset{\sim}{S}$ and $\underset{\sim}{\gamma}$, it can usually be extended to include multidimensional vectors [2.2]. We shall retain the vector notation in the following for illustrative purposes, noting that this extension is necessary for the validity of results in the multidimensional cases.

$$r(\underline{S},\gamma_s) = \int_{\Delta} P_3(\underline{\gamma}|\underline{S})\, C(\underline{S},\underline{\gamma})\, d\underline{\gamma} \tag{4.4}$$

where *

$$P_3(\underline{\gamma}|\underline{S}) = \int_{\Gamma} F_s(\underline{V})\, \delta(\underline{\gamma}-\gamma_s(\underline{V}))\, d\underline{V} \quad . \tag{4.5}$$

The latter is the probability (density) of making an estimate $\underline{\gamma}$ when the signal is $\underline{S}$, and is thus an error probability analogous to the error probabilities α, β in binary detection, cf. [3.7].

Next, the average risk for an a priori signal distribution $\sigma(\underline{S})$ is expressed as:

$$R(\sigma,\gamma_s) = \int_{\Omega}\int_{\Gamma} [\underline{S}-\gamma_s(\underline{V})]^2\, \sigma(\underline{S})\, F_s(\underline{V})\, d\underline{V} \quad , \tag{4.6}$$

for the squared-error cost function. Since

$$\sigma(\underline{S})\, F_s(\underline{V}) = f(\underline{V})\, P_1(\underline{S}|\underline{V}) \tag{4.7}$$

this may be rewritten

$$R(\sigma,\gamma_s) = \int_{\Gamma} d\underline{V}\, f(\underline{V}) \int_{\Omega} [\underline{S} - \gamma_s(\underline{V})]^2\, P_1(\underline{S}|\underline{V})\, d\underline{S} \quad . \tag{4.8}$$

The $\gamma_s(\underline{V})$ for which this is smallest is the Bayes estimator, denoted by $\gamma_\sigma(\underline{V})$. The second integral is a minimum for fixed $\underline{V}$ if $\gamma_s(\underline{V})$ is chosen as

$$\gamma_s(\underline{V})_{\min} \equiv \gamma_\sigma(\underline{V}) = \int_{\Omega} \underline{S}\, P_1(\underline{S}|\underline{V})\, d\underline{S} \tag{4.9}$$

or

$$\gamma_\sigma(\underline{V}) = \frac{\int_{\Omega} \underline{S}\, \sigma(\underline{S})\, F_s(\underline{V})\, d\underline{V}}{\int_{\Omega} \sigma(\underline{S})\, F_s(\underline{V})\, d\underline{S}} \, . \tag{4.10}$$

Thus, the Bayes estimator (for the cost functions of (4.6)) is the conditional expectation of $\underline{S}$ given $\underline{V}$.

The conditional risk (4.3) becomes the variance of the estimator when the cost function is squared-error and $\gamma_s(\underline{V})$ is unbiased for every $\underline{S}$, i.e., when

$$\int_{\Gamma} \gamma_s(\underline{V})\, F_s(\underline{V})\, d\underline{V} = \underline{S} \quad . \tag{4.11}$$

In this case the average risk (4.6) may be written as

$$R(\sigma,\gamma_s) = \int_{\Omega} \text{Var.}\ \gamma_s(\underline{V})\, \sigma(\underline{S})\, d\underline{S} \, . \tag{4.12}$$

Thus, the Bayes estimator, which minimizes $R(\sigma,\gamma_s)$ is a minimum variance estimator for every $\underline{S}$. However, the Bayes estimator (4.10) is unbiased only for certain distributions.

When signal and noise belong to ergodic processes, the average risk (4.6) may be written as

$$R(\sigma,\gamma_s) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \left\{ S(t) - \gamma_s\,[V(t)] \right\}^2 dt \, . \tag{4.13}$$

*Note that $P_1(\underline{S}|\underline{V})$, $P_2(\underline{S}|\underline{\gamma})$ and $P_3(\underline{\gamma}|\underline{S})$ are all different functions.

The conventional treatment of extraction uses the minimum value of this as an optimum criterion, just as the risk formulation does when the cost is squared-error. Usually, γ_s is required to be the output of a linear, physically realizable filter with V(t) at its input. Since (4.10) is not generally linear in V, optimum extractors under this constraint are not necessarily Bayes. We observe, however, that some of the ideas of risk theory are useful for such restricted classes of decision rules, even though the main theorems do not apply. That is, if we agree that only the class of linear estimators is to be considered, we may speak of the one with smallest average risk, the one for which the maximum conditional risk is smallest, the one with the property that no other is uniformly better, etc., settling the questions of existence and uniqueness in specific cases by construction.

Further detailed properties of Bayes and Minimax extractors, with results for reception problems of practical interest, are reserved for later presentation.

4.2. The Maximum Likelihood Estimator

A most useful method of obtaining estimates is furnished by the principle of maximum likelihood, which takes the $\underset{\sim}{S}(=\underset{\sim}{S}_{ML})$ that maximizes the likelihood function (i.e., $F_s(\underset{\sim}{V})$ regarded as a function of $\underset{\sim}{S}$ for fixed (i.e., given) $\underset{\sim}{V}$) as the best estimate of the actual $\underset{\sim}{S}$ present at the input.

Now, it is well known that when a sufficient statistic exists, the maximum likelihood estimator depends on it alone [4.3]. Its further significance is easily seen from (4.5). If (4.5) is written for $\underset{\sim}{\gamma}=\underset{\sim}{S}$, we have

$$P_3(\underset{\sim}{\gamma}=\underset{\sim}{S}\,|\,\underset{\sim}{S}) = \int_{\Gamma} F_s(\underset{\sim}{V})\ \delta(\underset{\sim}{S}-\underset{\sim}{\gamma}_s(\underset{\sim}{V}))\ d\underset{\sim}{V}\ , \tag{4.14}$$

which is therefore the probability of a correct decision when the signal is $\underset{\sim}{S}$. It is clearly largest if for each $\underset{\sim}{V}$ we choose $\underset{\sim}{\gamma}_s(\underset{\sim}{V})$ equal to $\underset{\sim}{S}_{ML}$. The conditional risk (4.4) is the sum of the products of the various error probabilities and their costs. Since the cost of a correct decision is always less than any other (by definition), the maximum likelihood estimator, by assigning the largest probabilities to the smallest costs, minimizes the risk if certain symmetries are present in $F_s(\underset{\sim}{V})$ as a function of $\underset{\sim}{S}$ and in the cost function $C(\underset{\sim}{S},\underset{\sim}{\gamma})$. Wald [4.4] shows that the maximum likelihood estimator of a one-dimensional parameter $\underset{\sim}{S} = S$ minimizes the Bayes risk when the cost depends only on the difference $|\gamma-S|$, when the *a priori* distribution $\sigma(S)$ is a constant, and when $F_s(\underset{\sim}{V}) = W(\underset{\sim}{V}-\underset{\sim}{S})$ is symmetric about $\underset{\sim}{S}_{ML}$.

By taking the average of each side of (4.14) with respect to an *a priori* signal distribution $\sigma(S)$, we obtain the average probability of a correct estimate as

$$\int_{\Omega} P_3(\underset{\sim}{S}\,|\,\underset{\sim}{S})\ \sigma(\underset{\sim}{S})\ d\underset{\sim}{S} = \int_{\Gamma} d\underset{\sim}{V}\ f(\underset{\sim}{V}) \int_{\Omega} P_1(\underset{\sim}{S}\,|\,\underset{\sim}{V})\ \delta(\underset{\sim}{S}-\underset{\sim}{\gamma}_s(\underset{\sim}{V}))\ d\underset{\sim}{S}\ . \tag{4.15}$$

The same reasoning that led to the maximization of (4.14) shows here that this is largest when for each $\underset{\sim}{V}$, $\underset{\sim}{\gamma}_s(\underset{\sim}{V})$ is chosen as the particular value of $\underset{\sim}{S}$ that makes the posterior probability $P_1(\underset{\sim}{S}\,|\,\underset{\sim}{V})$ largest. The Woodward and Davies receiver [1.19] presents $P_1(\underset{\sim}{S}\,|\,\underset{\sim}{V})$ as a function of $\underset{\sim}{S}$ at its output. We see, therefore, that if this is made into a "decision" system by taking the maximum value of the output as the estimate, the average probability of a correct decision is maximized. However, this procedure might be criticized, it ignores the possible importance of errors, i.e., failures of $\underset{\sim}{\gamma}$ to equal $\underset{\sim}{S}$ in one or more of its components, whose measure in the risk formulation is the cost function $C(\underset{\sim}{S},\underset{\sim}{\gamma})$.

The problem of estimating the amplitude of a small signal in additive noise furnishes an interesting application of the maximum likelihood method.

Assume that the form of the signal is known, that only one signal is possible, and that $\underset{\sim}{V}$ and $\underset{\sim}{S}$ are normalized: $\underset{\sim}{V} = \sqrt{\Psi}\,\underset{\sim}{\nu}$; $\underset{\sim}{S} = \sqrt{\Psi}\,a_o\underset{\sim}{s}$, with ψ the mean square noise amplitude. Expansion of $\log W(\underset{\sim}{\nu}-a_o\underset{\sim}{s})$ in powers of a_o, according to the procedure of Middleton [1.13, 1.21], yields the following, if we regard the terms in a_o and a_o^2 as an adequate representation of the distribution in the threshold case:

$$\begin{aligned}\log W_n(V-S_1;\ldots;V_n-S_n\,|\,a_o) = B_o^{(0)} &+ \left\{\sqrt{\Psi}\,\underset{\sim}{\tilde{\nu}}\,\underset{\sim}{B}^{(1)} + \psi\,\underset{\sim}{\tilde{\nu}}\,\underset{\sim}{B}^{(2)}\,\underset{\sim}{\nu} + \ldots\right\} \\ &- a_o\left\{\sqrt{\Psi}\,\underset{\sim}{\tilde{s}}\ \ \underset{\sim}{B}^{(1)} + 2\psi\,\underset{\sim}{\tilde{\nu}}\,\underset{\sim}{B}^{(2)}\,\underset{\sim}{s}\right\} \\ &+ a_o^2\left\{\psi\,\underset{\sim}{\tilde{s}}\,\underset{\sim}{B}^{(2)}\,\underset{\sim}{s}\right\} + \ \ (a_o^3)\ .\end{aligned} \tag{4.16}$$

Here $B^{(0)}$ is a scalar, $\underset{\sim}{B}^{(1)}$ a vector, and $\underset{\sim}{B}^{(2)}$ a square matrix, and all depend in general on the means,

variances and all higher moments of the noise distribution. The maximum likelihood estimate for this representation of the likelihood function is easily shown to be

$$a_{o_{ML}} \doteq \frac{\sqrt{\psi}\,\underset{\sim}{\hat{s}}\,\underset{\sim}{B}^{(1)} + 2\,\underset{\sim}{\hat{v}}\,\underset{\sim}{B}^{(2)}\,\underset{\sim}{s}}{2\,\underset{\sim}{\hat{s}}\,\underset{\sim}{B}^{(2)}\,\underset{\sim}{s}} \tag{4.17}$$

in this approximation. Substitution of this back into (4.16) gives

$$W(\underset{\sim}{V}-\underset{\sim}{S}) \doteq \left\{\exp\left[-2\psi a_{o_{ML}} a_o\,\underset{\sim}{\hat{s}}\,\underset{\sim}{B}^{(2)}\,\underset{\sim}{s} + \psi a_o^2\,\underset{\sim}{\hat{s}}\,\underset{\sim}{B}^{(2)}\,\underset{\sim}{s}\right]\right\} \exp\left[B^{(0)} + \sqrt{\psi}\,\underset{\sim}{\hat{v}}\,\underset{\sim}{B}^{(1)} + \psi\,\underset{\sim}{\hat{v}}\,\underset{\sim}{B}^{(2)}\,\underset{\sim}{v}\right] \quad . \tag{4.18}$$

Since the distribution factors into one term involving the statistic $a_{o_{ML}}$ and the parameter a_o, and another independent of the parameter, the estimator $a_{o_{ML}}$ is a sufficient statistic, [4.3] cf Sec. 2.5. When the noise has a Gaussian structure

$$\underset{\sim}{B}^{(1)} = \sqrt{\psi}\,\underset{\sim}{K}^{-1}\,\underset{\sim}{\bar{v}}$$

$$\underset{\sim}{B}^{(2)} = -\frac{1}{2}\,\underset{\sim}{K}^{-1} \quad , \tag{4.19}$$

where $\underset{\sim}{K}$ is the variance matrix and the result (4.17), (4.18) is exact.

4.3. System Evaluation and Comparison; Analogies between Detection and Extraction

Here, as for detection, systems may be evaluated and compared in risk terms. When a system performs both detection and extraction, one may consider the corresponding two risks as components of the total risk associated with "complete" reception of the signal. Thus, the extraction risk of any given system may be calculated from (4.2) with the system function in place of $\gamma_s(\underset{\sim}{V})$.

Figure 5 briefly recapitulates the risk formulations for detection and extraction. For each, the conditional risk is the sum of the various error probabilities for a given $\underset{\sim}{S}$ $(P_3(\underset{\sim}{\gamma}|\underset{\sim}{S}))$, weighted according to error cost $C(\underset{\sim}{S}|\underset{\sim}{\gamma})$. The average risk is the expected value of this in view of the prior signal distribution $\sigma(S)$. The essential difference between the two appears in the calculation of error probabilities. For extraction this amounts to a simple "folding" of the distribution $F(\underset{\sim}{V}|\underset{\sim}{S})$, i.e., a selection of all the $\underset{\sim}{V}$'s that lead to a given decision $\underset{\sim}{\gamma}$ and a summation of their probabilities of occurrence. For detection there is seen to be a similar folding operation, with the likelihood ratio $\mathcal{L}(\underset{\sim}{V})$ taking the place of the estimator $\gamma_s(\underset{\sim}{V})$, followed by an additional summation over all values of $\mathcal{L}(V)$ above or below the threshold K. Thus detection essentially involves an extraction type of operation which maps observation space Γ onto the real line (domain of y) with subsequent division of this domain into two parts at the threshold value K. (See remarks in Sec. 1 on this connection between detection and extraction.)

5. Connection Between Information Loss and Risk

5.0. Introduction

In this section we show how, as a special case of decision theory, the Shannon measure of information loss may be used as a criterion of performance for detection and extraction systems. Systems that minimize information loss are described, and some of the relations between the minimum information loss criterion and the minimum risk criterion are pointed out.

As mentioned in Sec. 1.1, the general formulation from the point of view of information loss and risk are the same except that the cost function $C(\underset{\sim}{S},\underset{\sim}{\gamma})$ of the latter is replaced by the uncertainty $-\log P_2(\underset{\sim}{S}|\underset{\sim}{\gamma})$. The average information loss, or equivocation, is thus given by

$$H(\sigma,\delta) = -\int_{\Omega}\int_{\Gamma}\int_{\Delta} \log P_2(\underset{\sim}{S}|\underset{\sim}{\gamma})\,\sigma(\underset{\sim}{S})\,F_s(\underset{\sim}{V})\,\delta(\underset{\sim}{\gamma}|\underset{\sim}{V})\,d\underset{\sim}{S}\,d\underset{\sim}{V}\,d\underset{\sim}{\gamma} \quad , \tag{5.1}$$

where again $\delta(\underset{\sim}{\gamma}|\underset{\sim}{V})$ is the decision rule, assumed to be nonrandom. Since the value of $P_2(\underset{\sim}{S}|\underset{\sim}{\gamma})$ for given $\underset{\sim}{S}$ and $\underset{\sim}{\gamma}$ depends on the decision rule in use, while that of $C(\underset{\sim}{S},\underset{\sim}{\gamma})$ does not, the decision rule that minimizes information loss is harder to find than the one that minimizes risk.

5.1. The Information Loss Criterion for Detection.

To specialize the expression (5.1) for (binary) detection, we assume first that the (nonrandomized) decision rule divides observation space Γ into two regions, here denoted as Γ_0 and Γ_1, and that the decision γ_0 (noise alone) is made when $\underset{\sim}{V} \in \Gamma_0$ ($\in$ = "lies in" or "belongs to"), and the decision γ_1, (signal and noise) when $\underset{\sim}{V} \in \Gamma_1$. Thus we write

$$\delta(\gamma_i | \underset{\sim}{V} \in \Gamma_j) = \delta_{ij} \quad , \quad i,j = 0,1, \tag{5.2}$$

where $\delta_{ij} = 0$ $(i \neq j)$ or $1(i=j)$. With this, (5.1) becomes

$$H(\sigma,\delta) = - \int_{\Omega} \sigma(\underset{\sim}{S}) \left\{ P_3(\gamma_0|\underset{\sim}{S}) \log P_2(\underset{\sim}{S}|\gamma_0) + P_3(\gamma_1|\underset{\sim}{S}) \log P_2(\underset{\sim}{S}|\gamma_1) \right\} d\underset{\sim}{S}. \tag{5.3}$$

The equivocation for any given binary detection system may be calculated from (5.3) and used to judge its performance [5.1]. As an example, let us suppose for the moment that either $\underset{\sim}{S} = \underset{\sim}{S}_0 = 0$ or $\underset{\sim}{S} = \underset{\sim}{S}_1 \neq 0$ can be present at the input with a priori probabilities $\sigma(\underset{\sim}{S}_0) = q$ and $\sigma(\underset{\sim}{S}_1) = p$ (simple alternative case). Then $\sigma(\underset{\sim}{S})$ equals $\sigma(S_0)\ \delta(\underset{\sim}{S}-\underset{\sim}{S}_0) + \sigma(S_1)\ \delta(\underset{\sim}{S}-\underset{\sim}{S}_1)$, and (5.3) becomes

$$H(\sigma,\delta) = -\sum_{i,j} \sigma(\underset{\sim}{S}_i)\ P_3(\gamma_j|\underset{\sim}{S}_i) \log P_2(\underset{\sim}{S}_i|\gamma_j) \quad , \quad i,j = 0,1. \tag{5.4}$$

P_2 and P_3 may readily be expressed in terms of the error probabilities α and β, as follows:

$$P_3(\gamma_0|\underset{\sim}{S}_0) = 1 - \alpha \quad , \quad P_3(\gamma_0|\underset{\sim}{S}_1) = \beta \quad ,$$

$$P_3(\gamma_1|\underset{\sim}{S}_0) = \alpha \quad , \quad P_3(\gamma_1|\underset{\sim}{S}_1) = 1 - \beta \quad , \tag{5.5}$$

$$P_2(\underset{\sim}{S}_0|\gamma_0) = \frac{q(1-\alpha)}{q(1-\alpha)+p\beta} \quad , \quad P_2(\underset{\sim}{S}_0|\gamma_1) = \frac{q\alpha}{q\alpha+p(1-\beta)} \quad ,$$

$$P_2(\underset{\sim}{S}_1|\gamma_0) = \frac{p\beta}{q(1-\alpha)+p\beta} \quad , \quad P_2(\underset{\sim}{S}_1|\gamma_1) = \frac{p(1-\beta)}{q\alpha+p(1-\beta)} \quad . \tag{5.6}$$

Figure 6 shows how the equivocation of Bayes tests depends on the cost ratio K and $a_o^2\sqrt{n}$, for $\mu = 1$ (and Rayleigh statistics). These curves are calculated, with the help of Figure 3 to find α and β, and by evaluating (5.5) and (5.6) followed by substitution into (5.4). In these circumstances we see that the Ideal Observer (K=1) loses the least information for a given signal amplitude and integration time. As the cost ratio K is varied for fixed $a_o^2\sqrt{n}$, the minimum at K = 1 appears broad; in fact, the information loss for any fixed $a_o^2\sqrt{n}$ does not vary by more than 0.2 bit as K is changed from 1/16 to 16. These curves may be used to define a minimum detectable signal (for threshold detection) in the same way as the Bayes risk curves are used. Accordingly, if 0.2 bit is taken as the largest allowable loss, the minimum value of $a_o^2\sqrt{n}$ is 3.75 for K = 1 and 4.25 for K = 16, 1/16. For fixed sample size n this amounts to a difference of only 0.54 db in the amplitude of the minimum detectable signal. Correspondingly, for fixed signal amplitudes, the change in integration time is only $\sqrt{4.25/3.75} = 1.06$, or 6 per cent. For very small signals the equivocation approaches 1 bit, so that the system does no better than one who guesses on the basis of the a priori probabilities. For large signals or integration times, on the other hand, the equivocation approaches zero, corresponding to the increasing certainty of a correct decision.

Let us now return to the general expression for equivocation, (5.3), and seek to minimize it by choice of the decision rule, i.e., here by choice of the boundary between regions Γ_0 and Γ_1, without specific assumptions as yet about the prior signal distribution $\sigma(S)$. The probability functions P_2 and P_3 may be written

$$P_2(\underset{\sim}{S}|\gamma_i) = \int_{\Gamma_i} P_1(\underset{\sim}{S}|\underset{\sim}{V})\ P_4(\underset{\sim}{V}|\gamma_i)\ d\underset{\sim}{V} \quad , \tag{5.7}$$

$$P_3(\gamma_i|\underset{\sim}{S}) = \int_{\Gamma_i} F_s(\underset{\sim}{V})\ d\underset{\sim}{V} \quad . \tag{5.8}$$

Here $P_4(\underset{\sim}{V}|\gamma_i)$ is the probability that $\underset{\sim}{V}$ was responsible for the decision γ_i. Since the decision rule is

nonrandomized, every $\underset{\sim}{V}$ in Γ_i leads to the decision γ_i with the probability 1, and the $\underset{\sim}{V}$'s outside of Γ_i cannot lead to γ_i at all. Thus $P_4(\underset{\sim}{V}|\gamma_i)$ is constant throughout Γ_i, equal in fact to the reciprocal of the "volume" of Γ_i, since P_4 must be properly normalized. Thus,

$$P_4(\underset{\sim}{V}|\gamma_i) = 1/\Gamma_i \quad , \tag{5.9}$$

$$P_2(\underset{\sim}{S}|\gamma_i) = \frac{1}{\Gamma_i}\int_{\Gamma_i} P_1(\underset{\sim}{S}|\underset{\sim}{V})\, d\underset{\sim}{V} \quad , \tag{5.10}$$

where we have let Γ_i stand here for the volume of the region as well as for the domain of $\underset{\sim}{V}$ included within the volume. Now denoting by $\underset{\sim}{V}'$ a point on the boundary between Γ_o and Γ_1, and letting Γ_o be increased to $\Gamma_o' + d\Gamma_o$ by change of $\underset{\sim}{V}'$ to $\underset{\sim}{V}' + d\underset{\sim}{V}'$, we find for the derivatives involved in the minimization:

$$\frac{\partial}{\partial \underset{\sim}{V}'} P_3(\gamma_o|\underset{\sim}{S}) = -\frac{\partial}{\partial \underset{\sim}{V}'} P_3(\gamma_1|\underset{\sim}{S}) = F_s(\underset{\sim}{V}') \quad , \tag{5.11a}$$

$$\frac{\partial}{\partial \underset{\sim}{V}'} \log P_2(\underset{\sim}{S}|\gamma_o) = \frac{P_1(\underset{\sim}{S}|\underset{\sim}{V}')}{\Gamma_o P_2(\underset{\sim}{S}|\gamma_o)} - \frac{1}{\Gamma_o} \quad , \tag{5.11b}$$

$$\frac{\partial}{\partial \underset{\sim}{V}'} \log P_2(\underset{\sim}{S}|\gamma_1) = \frac{-P_1(\underset{\sim}{S}|\underset{\sim}{V}')}{\Gamma_1 P_2(\underset{\sim}{S}|\gamma_1)} + \frac{1}{\Gamma_1} \quad . \tag{5.11c}$$

The derivative of the bracket in the integrand of (5.3) then becomes

$$\frac{P_1(\underset{\sim}{S}|\underset{\sim}{V}')}{\sigma(\underset{\sim}{S})}\left\{\frac{P_5(\gamma_o)}{\Gamma_o} - \frac{P_5(\gamma_1)}{\Gamma_1}\right\} + \left\{\frac{P_3(\gamma_1|\underset{\sim}{S})}{\Gamma_1} - \frac{P_3(\gamma_o|\underset{\sim}{S})}{\Gamma_o}\right\} + F_s(\underset{\sim}{V}')\log\frac{P_2(\underset{\sim}{S}|\gamma_o)}{P_2(\underset{\sim}{S}|\gamma_1)} \quad , \tag{5.12}$$

where we have used the relation

$$P_2(\underset{\sim}{S}|\gamma_i) = \frac{\sigma(\underset{\sim}{S})}{P_5(\gamma_i)} P_3(\gamma_i|\underset{\sim}{S}) \quad . \tag{5.13}$$

Here $P_5(\gamma_i)$ is the (total) probability of making decision γ_i, given by

$$P_5(\gamma_i) = \int_\Omega P_3(\gamma_i|\underset{\sim}{S})\,\sigma(\underset{\sim}{S})\, d\underset{\sim}{S} \quad . \tag{5.14}$$

The integration over Ω indicated in (5.3) causes the first four terms in (5.12) to cancel, leaving finally

$$\frac{\partial H(\sigma,\delta)}{\partial \underset{\sim}{V}'} = -\int_\Omega \sigma(\underset{\sim}{S})\, F_s(\underset{\sim}{V}')\log\frac{P_2(\underset{\sim}{S}|\gamma_o)}{P_2(\underset{\sim}{S}|\gamma_1)}\, d\underset{\sim}{S} = 0 \quad . \tag{5.15}$$

This is the condition for minimum (or maximum) equivocation, i.e., the boundary between Γ_o and Γ_1 must be such that this relation is satisfied.

The results for the one-sided and simple alternative tests are obtained from (5.15) by suitable specialization of $\sigma(S)$. For the one-sided alternative we take as before $\sigma(\underset{\sim}{S}) = q\,\delta(\underset{\sim}{S}-0) + w(\underset{\sim}{S})$, which yields (with $F_s(\underset{\sim}{V})$ now replaced by $W(\underset{\sim}{V}-\underset{\sim}{S})$ in the case of additive noise):

$$\frac{\langle W(\underset{\sim}{V}'-\underset{\sim}{S})\rangle_{(\underset{\sim}{S}|\gamma)}}{q\,W(\underset{\sim}{V}')} = -\log\frac{P_2(\underset{\sim}{S}=0|\gamma_o)}{P_2(\underset{\sim}{S}=0|\gamma_1)} \quad , \tag{5.16}$$

where

$$\langle W(\underset{\sim}{V}'-\underset{\sim}{S})\rangle_{(\underset{\sim}{S}|\gamma)} = \int_\Omega W(\underset{\sim}{V}-\underset{\sim}{S})\, w(\underset{\sim}{S})\log\frac{P_2(\underset{\sim}{S}|\gamma_o)}{P_2(\underset{\sim}{S}|\gamma_1)}\, d\underset{\sim}{S} \quad . \tag{5.17}$$

Equations (5.16) and (5.17) show that the optimum (or extremal) division of observation space is achieved here by a generalized likelihood ratio test in which $W(\underset{\sim}{V}'-\underset{\sim}{S})$ is averaged with respect to a distribution $w(\underset{\sim}{S})\log[P_2(\underset{\sim}{S}|\gamma_o)/P_2(\underset{\sim}{S}|\gamma_1)]$, which itself depends on the optimum (or extremal) division.

For the simple alternative case we have $\sigma(\underset{\sim}{S}) = q\,\delta(\underset{\sim}{S}-0) + p\,\delta(\underset{\sim}{S}-S_1)$, so that (5.15) becomes*

$$\frac{pW(\underset{\sim}{Y}'-\underset{\sim}{S})}{qW(\underset{\sim}{Y}')} = K_H \tag{5.18}$$

where

$$K_H = \frac{\log z_o}{\log z_1} \tag{5.19}$$

and

$$z_o = \frac{P_2(\underset{\sim}{S}=o|\gamma_o)}{P_2(\underset{\sim}{S}=o|\gamma_1)}\,, \tag{5.20a}$$

$$z_1 = \frac{P_2(\underset{\sim}{S}_1|\gamma_o)}{P_2(\underset{\sim}{S}_1|\gamma_1)}\,. \tag{5.20b}$$

The values of $\underset{\sim}{Y}'$ satisfying equation (5.18) define the extremum boundary between Γ_o and Γ_1, i.e., if "noise alone" is decided whenever $\underset{\sim}{Y}$ falls within Γ_o and "signal and noise" when $\underset{\sim}{Y}$ falls within Γ_1, the information loss is a maximum or a minimum for this division.

Note that Equation (5.18) defines a likelihood ratio test of the same type as the Bayes test for the corresponding problem in the risk formulation. Thus tests that minimize information loss (when they exist) belong to the Bayes class and are equivalent to minimum average risk tests with special cost assumptions. It is therefore possible for a system to be optimum simultaneously from the standpoints of both risk and information loss.

To show that there are solutions of (5.18) that minimize information loss, we may use (5.6) to express z_o and z_1 of (5.20) in terms of p, q, α and β. With $\mu = p/q$ the result is

$$z_o = \mu\left(\frac{1-\alpha}{\alpha}\right)\left(\frac{1+\alpha/\mu-\beta}{1-\alpha+\beta\mu}\right)\,, \tag{5.21a}$$

$$z_1 = \frac{1}{\mu}\left(\frac{1-\beta}{\beta}\right)\left(\frac{1-\alpha+\beta\mu}{1+\alpha/\mu-\beta}\right)\,. \tag{5.21b}$$

It may readily be shown that for $\alpha + \beta < 1$ (the case of ordinary interest), z_o and z_1 are both greater than unity, so that the likelihood threshold K_H is always positive. [We note that interchange of α and β and of p and q interchanges z_o and z_1, thus inverting K_H.]

The universal curve of Figure 7 shows the relation between α, β, and K_H for $\mu = 1$, exhibiting this symmetry. We see that the existence of a minimum information loss test depends on whether the statistics of the problem admit a likelihood ratio test with α and β related to the threshold K_H, as shown in Figure 7. If, for example, Figure 7 is superimposed on Figure 3, which gives the characteristics for Rayleigh statistics, we observe that there are no combinations of α, β and K (for $\mu = 1$) that fit both at once, except for those alone the line $\alpha = \beta$, $K = K_H = 1$. Figure 6 shows that the information loss is indeed a minimum for this value of K, so that for Rayleigh statistics the Ideal Observer (who takes K = 1) minimizes information loss and risk simultaneously (when $\mu = 1$).

On the other hand, we note that since the Neyman-Pearson observer does not generally minimize information loss, as the example of Figure 6 shows, and yet is optimum in a risk sense when fixed false alarm time is important, we may conclude that it is not always necessarily desirable for a system to minimize information loss. Detailed control of decision error may be more important in some cases.

Thus, although tests that minimize information loss are likelihood ratio tests, and therefore form a subclass of the Bayes tests, they exist under much less general conditions than do the minimum risk tests. That is, the latter exist for any given cost ratio and μ, while the former exist only for certain cost ratios and μ's, depending on the statistics. Determination of the broad conditions under which the information loss extremum exists and, moreover, is a minimum, awaits further investigation.

*W.M. Siebert and R.M. Lerner (M.I.T.) in a recent unpublished memorandum have independently obtained a similar result by a different method.

5.2. The Information Loss Criterion for Extraction.

To specialize (5.1) for extraction we first assume that the decision rule is nonrandomized and use (4.1) to obtain

$$H(\sigma,\delta) = -\int d\underset{\sim}{V}\, f(\underset{\sim}{V}) \int_{\Omega} P_1(\underset{\sim}{S}|\underset{\sim}{V}) \log P_2[\underset{\sim}{S}|\gamma(\underset{\sim}{V})]\, d\underset{\sim}{S} \quad . \tag{5.22}$$

This expression may be minimized by choosing $\underset{\sim}{\gamma}$ to minimize the second integral for arbitrary fixed $\underset{\sim}{V}$. As before, P_2 may be expressed as

$$P_2(\underset{\sim}{S}|\underset{\sim}{\gamma}) = \int_{\Gamma_\gamma} P_1(\underset{\sim}{S}|\underset{\sim}{V})\, P_4(\underset{\sim}{V}|\underset{\sim}{\gamma})\, d\underset{\sim}{V}, \tag{5.23}$$

where Γ_γ denotes the domain of all $\underset{\sim}{V}$'s that lead to the decision γ. By the argument used previously (cf. (5.9)), P_4 is constant over the region Γ_γ and zero outside so that

$$P_4(\underset{\sim}{V}|\underset{\sim}{\gamma}) = \frac{1}{N(\underset{\sim}{\gamma})} \quad , \quad \text{where } N(\underset{\sim}{\gamma}) = \int_{\Gamma_\gamma} d\underset{\sim}{V} \quad . \tag{5.24}$$

Thus, (5.23) becomes

$$P_2(\underset{\sim}{S}|\underset{\sim}{\gamma}) = \frac{M_S(\underset{\sim}{\gamma})}{N(\underset{\sim}{\gamma})} \quad , \quad \text{where } M_S(\underset{\sim}{\gamma}) = \int_{\Gamma_\gamma} P_1(\underset{\sim}{S}|\underset{\sim}{V})\, d\underset{\sim}{V} \quad . \tag{5.25}$$

Differentiating (5.22) with respect to $\underset{\sim}{\gamma}$ we obtain the following condition for an information loss extremum:

$$\int_{\Omega} P_1(\underset{\sim}{S}|\underset{\sim}{V}) \left\{ \frac{M_S'(\underset{\sim}{\gamma})}{M_S(\underset{\sim}{\gamma})} - \frac{N'(\underset{\sim}{\gamma})}{N(\underset{\sim}{\gamma})} \right\} d\underset{\sim}{S} = 0, \tag{5.26}$$

where the primes denote differentiation with respect to $\underset{\sim}{\gamma}$. In view of the definitions of (5.24) and (5.25), this states the requirement to be fulfilled by Γ_γ. The optimum (or extremal) rule for obtaining $\underset{\sim}{\gamma}$ from $\underset{\sim}{V}$ must be such that it produces Γ_γ's with the properties implied by (5.26).

We note immediately that (5.26) is satisfied if $\underset{\sim}{\gamma}$ is a sufficient statistic, i.e., if $P_1(\underset{\sim}{S}|\underset{\sim}{V}) = P_2(\underset{\sim}{S}|\underset{\sim}{\gamma})$, $\underset{\sim}{V} \in \Gamma_\gamma$; (see Sec. 2.5). For in that case we have

$$M_S(\underset{\sim}{\gamma}) = P_2(\underset{\sim}{S}|\underset{\sim}{\gamma})\, N(\underset{\sim}{\gamma}) \quad , \tag{5.27}$$

and the bracket in (5.26) becomes

$$\frac{1}{P_1(\underset{\sim}{S}|\underset{\sim}{V})} \frac{\partial}{\partial \underset{\sim}{\gamma}} P_2(\underset{\sim}{S}|\underset{\sim}{\gamma}) \quad , \tag{5.28}$$

which satisfies (5.26) identically. The sufficiency condition results alternatively, if (5.22) is minimized directly with respect to unconstrained variation of the function $P_2(\underset{\sim}{S}|\underset{\sim}{\gamma})$. When the distribution $F_s(\underset{\sim}{V})$ does not admit a sufficient statistic, however, (5.26) gives the condition for an extremum. Specific tests that fulfill this condition remain to be investigated.

6. Summary Remarks

This paper has attempted to demonstrate some of the advantages of regarding reception systems in communication as systems for making statistical decisions with minimum risk or information loss. This approach, closely related to game theory, assumes a loss function is assigned at the outset to each possible decision error, in some cases depending as well on the choice of decision rule. This emphasizes the obvious but often overlooked fact that the criterion of best performance is not absolute but arbitrary, its merit depending on how well the loss function chosen reflects the constraints on the problem and the design objectives.

Formulation of the detection and extraction problems within this framework exhibits their close

relationship. Detailed analysis reveals the nature of optimum detection and extraction systems from each point of view, showing how previously used criteria appear as special cases of the more general formulation. Methods for comparing actual and ideal systems are also briefly outlined, whereby performance sacrificed by compromise in design may, at least in principle, be found. The central rôle of a priori statistical knowledge in practical system design is stressed throughout, and Minimax methods for handling incomplete knowledge of this type are discussed and illustrated.

The new feature of this work is the adaptation of recent advances in the theory of statistical inference to practical communication problems. The general formulation appears broad enough to form the basis of attack on many special problems of interest.

Bibliography

[1.1]. N. Wiener, "The Extrapolation, Interpolation, and Smoothing of Stationary Time Series," John Wiley (New York) 1949.

[1.2]. H.E. Singleton, Theory of Nonlinear Transducers, Tech. Report No. 160, Res. Lab. Electronics (M.I.T.) August (1950).

[1.3]. L.A. Zadeh and J.R. Ragazzini, An Extension of Wiener's Theory of Prediction , J. Appl. Phys. 21, 645 (1950).

[1.4]. R.C. Booton, Jr., An Optimization Theory for Time-varying Linear Systems with Non-Stationary Statistical Inputs, Proc. I.R.E. 40, 977 (1952).

[1.5]. R.C. Davis, On the Theory of Prediction of Nonstationary Stochastic Processes, J. Appl. Phys. 23, 1047 (1952).

[1.6]. D.O. North, Analysis of Factors which Determine Signal-to-Noise Discrimination in Pulsed Carrier Systems, R.C.A. Report PTR-6C (June, 1943).

[1.7]. J. H. Van Vleck and D. Middleton, A Theoretical Comparison of the Visual, Aural, and Meter Reception of Pulsed Signals in the Presence of Noise, J. App. Phys. 17, 940 (1946).

[1.8]. H. Den Hartog and F.A. Muller, Optimum Instrument Response for Discrimination against Spontaneous Fluctuations, Physica 13, 571 (1947).

[1.9]. B.M. Dwork, Detection of a Pulse Superimposed on Fluctuation Noise, Proc. I.R.E. 38, 771 (1950).

[1.10]. T.S. George, Fluctuations of Ground Clutter Return in Airborne Radar Equipment, J.I.E.E. 99, (IV) 92 (1952).

[1.11]. H. Urkowitz, Filter for Detection of Small Radar Signals in Clutter, J. Appl. Phys. 24, 1024 (1953).

[1.12]. L. Zadeh, Optimum Nonlinear Filters for the Extraction and Detection of Signals, J. Appl. Phys. 24, 396 (1953).

[1.13]. D. Middleton, The Statistical Theory of Signal Detection, Trans. Prof. Group Info. Theory PGIT-3, March (1954).

[1.14]. U. Grenander, Stochastic Processes and Statistical Inference, Arkiv. Mat. (Stockholm) 1 and 3, 197 and 277 (1950).

[1.15]. J.L. Lawson and G.E. Uhlenbeck, "Threshold Signals," McGraw-Hill (New York) 1950, Vol. 24, M.I.T. Rad. Lab. Series; Sec. (7.5).

[1.16]. H. Hanse, Doctoral Dissertation (M.I.T.), Jan. 1951.

[1.17]. E. Reich and P. Swerling, Jr., Detection of a Sine Wave in Gaussian Noise, J. Appl. Phys. 24, 289 (1953).

[1.18]. D. Middleton, Statistical Criteria for the Detection of Pulsed Carriers in Noise, I, II, J. Appl. Phys. 24, 371, 379 (1953).

[1.19]. P.M. Woodward and I. L. Davies, Information Theory and Inverse Probability in Telecommunications, Proc. I.E.E. 99 (III) 37 (1952).

[1.20]. I. L. Davies, On Determining the Presence of Signals in Noise, ibid. p. 45.

[1.21]. D. Middleton, Statistical Theory of Reception I: Optimum Detection of Signals in Noise, paper submitted to J. Appl. Phys. June, 1954.

[1.22]. J.L. Hodges and E.L. Lehmann, Some Problems in Minimax Point Estimation. Ann. Math. Stat. 21, 182 (1950).

[1.23]. Wassily Hoeffding, Optimum Nonparametric Tests. Proc. 2nd Berkely Symp. U. Cal. Press. p. 83 (1950).

[1.24]. E.L. Lehmann and C. Stein, On the Theory of Some Nonparametric Hypotheses. Ann. Math. Stat. 20, 28 (1949).

[1.25]. D. Middleton, The Statistical Theory of Detection I: Optimum Detection of Signals in Noise. M.I.T. Lincoln Lab. Technical Report No. 35. Nov. 1953.

[1.26]. A.J.F. Siegert, Passage of Stationary Processes Through Linear and Nonlinear Devices. Trans. I.R.E. PGIT-3,4, (1954).

[1.27]. R.C. Davis, The Detectability of Random Signals in the Presence of Noise. Trans. I.R.E. PGIT-3,52 (1954); also J. Appl. Phys. 25, 76 (1954)

[2.1]. A. Wald, "Statistical Decision Functions," John Wiley (New York) 1950.

[2.2]. D. Blackwell and M.A. Girshick, "Theory of Games and Statistical Decisions," John Wiley (New York) (1954).

[2.3]. E.W. Sampson, Fundamental Natural Concepts of Information Theory, AFCRC Report No. E 5079, Oct. 1951, Sec. 14.

[2.4]. C.E. Shannon, Mathematical Theory of Communication , Bell Sys. T. J. 27, 379, 623 (1948).

[2.5]. See, for example, J. Neyman, "Lectures and Conferences on Math. Stat. and Probability," 2nd Ed. Graduate School, U.S.D.A. Washington, 1950, p. 194.

[2.6]. A. Wald, Ref. [2.1], Sec. (1.6).

[2.7]. J.L. Hodges and E.L. Lehmann, The Use of Previous Experiences in Reaching Statistical Decisions. Ann. Math. Stat. 23, 396 (1952).

[2.8]. A. Wald, Ref. (2.1), Theorem (3.20) and ensuing remarks.

[2.9]. J. Kiefer, Am. Math. Stat. 24, 71 (1953) shows that Wald's restriction of D to the class of decision functions for which $r(\underline{S},\delta)$ is a bounded function of $\underline{S}$ is unnecessary.

[2.10]. A. Wald, op. cit. ref. [2.1]. Wald proved the theorem under less restrictive conditions, but these, (A) - (D), are sufficient for our applications. His assumptions (3.1)-(3.3), Chapter 3 are covered by (A) and (B), (3.5), (3.6) by (C), and (3.4), (3.7) by (D).

[2.11]. A. Wald, ref. [2.1]. Theorems (3.5), (3.7), (3.9), (3.14).

[2.12]. For reference to the original papers of Fisher, see H. Cramér, "Mathematical Methods of Statistics," Princeton (1947).

[2.13]. P.R. Halmos and L.J. Savage, Applications of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics. Ann. Math. Stat. 20, 225 (1949).

[2.14]. R.A. Fisher, Proc. Camb. Phil. Soc. 22, 700 (1925). For useful discussions of the properties of Fisher's information measure, see E.J.G. Pitman, Proc. Camb. Phil. Soc. 32, 567 (1936) and J. L. Hodges and E.L. Lehmann, Proc. Berkeley Symposium, U. Cal. Press, 1951. J.L. Doob, Trans. Am. Math. Soc. 39, 410 (1936) discusses another information measure with similar properties.

[2.15]. P.M. Woodward, Probability and Information Theory, with Applications to Radar, McGraw-Hill (1953); Sec. (3.7) gives this result without mention of sufficient statistics.

[2.16]. S. Kullback and R.A. Leibler, Ann. Math. Stat. 22, 79 (1951), in discussing hypothesis testing, use the logarithm of the likelihood ratio as a measure of the information contained in an observation, for discrimination between the two hypotheses, showing that its average value is positive semidefinite and invariant under a sufficient transformation. G.W. Preston, J. Appl. Phys. 24, 841 (1953) uses the result (Eq. 2.16), without proof, to define optimum extraction.

[3.1]. Kullback and Leibler, ref. [2.16]. The constant of proportionality involves the elements of Fisher's information matrix.

[3.2]. D. Middleton, ref. [1.18]. The relations (3.22), (3.23) here replace Eqs. (4.18), (4.19) of the reference, when one additional term is used in the approximation of $I_o(R\, a_o(2/\psi)^{1/2})$.

[3.3]. D. Middleton, Further Remarks on the Nature of the Statistical Observer, J. Appl. Phys. 25 127 (1954); also, D. Middleton, W.W. Peterson, and P.T. Birdsall, J. Appl. Phys. 25, 128 (1954).

[4.1]. L.A. Zadeh, General Filters for Separation of Signal and Noise. (Paper presented at Brooklyn Symposium on Information Networks, April 1954) discusses similar questions from a somewhat related viewpoint.

[4.2]. J.L. Hodges and E.L. Lehmann, Some Problems in Minimax Point Estimation, Ann. Math. Stat. 21, 182 (1950).

[4.3]. H. Cramér, Mathematical Methods of Statistics, Princeton (1947) 33.2, pg. 499.

[4.4]. A. Wald, Contributions to the Theory of Statistical Estimations and Testing Hypotheses. Ann. Math. Stat. 10, 299 (1939).

[5.1]. D. Middleton, Information Loss Attending the Decision Operation in Detection. J. Appl. Phys. 25, 127 (1954).

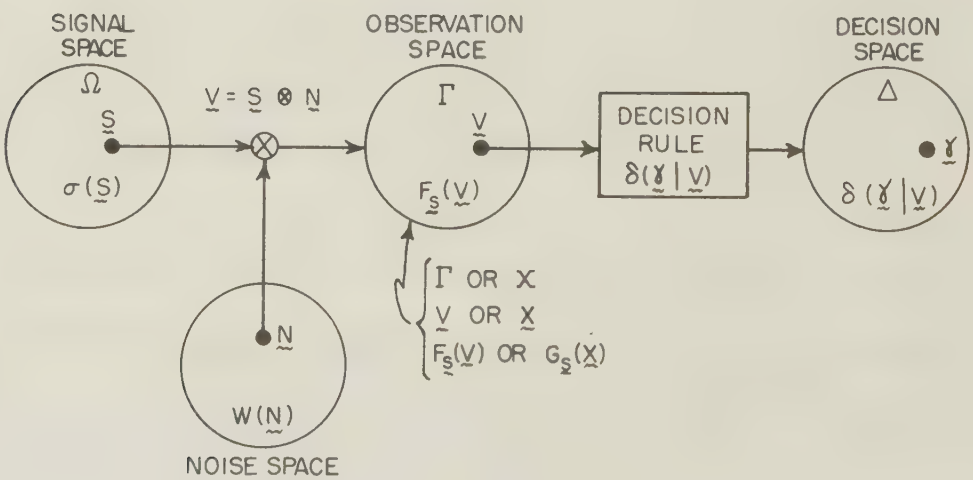


FIG. I THE DECISION SITUATION

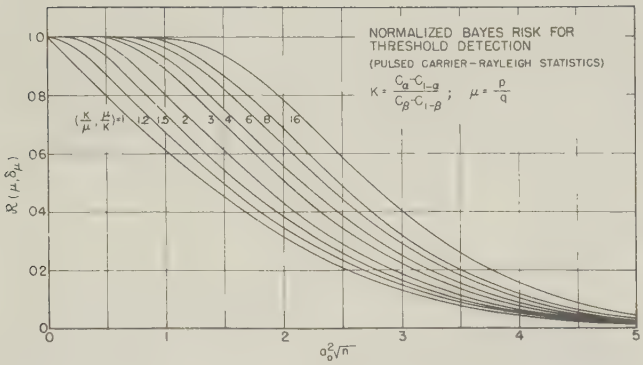


Fig. 2

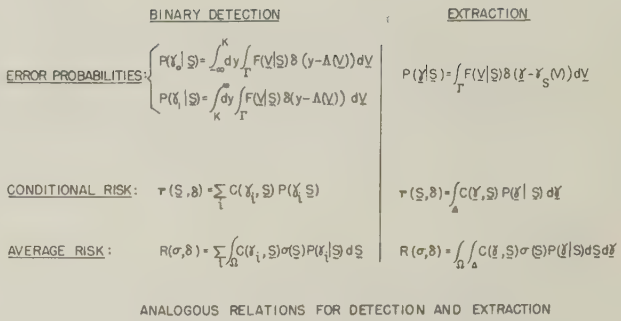


Fig. 5

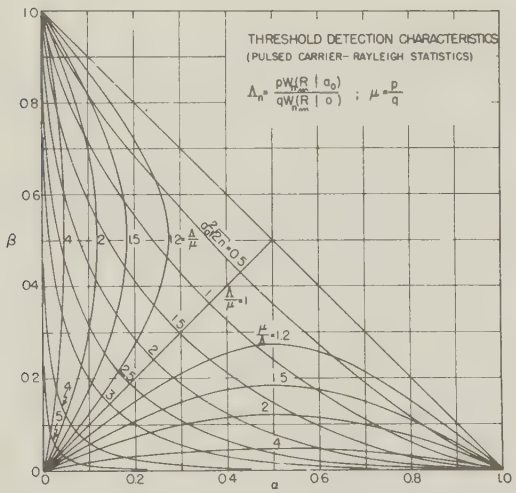


Fig. 3

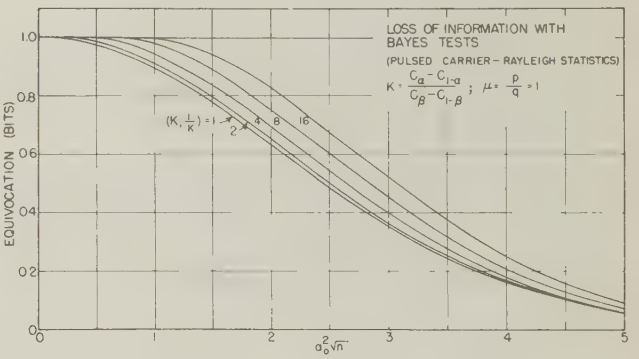


Fig. 6

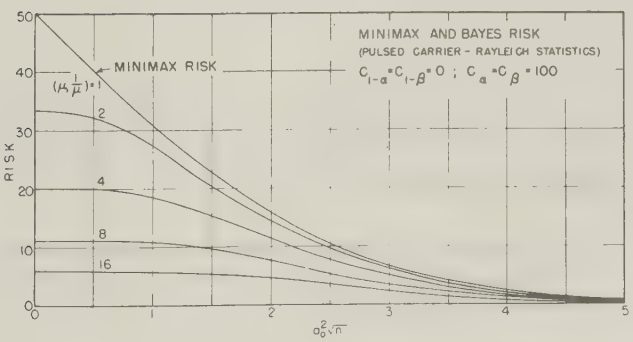


Fig. 4

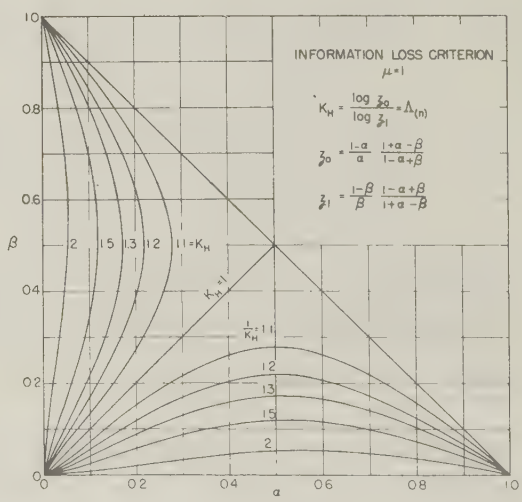


Fig. 7

A NON-LINEAR PREDICTION THEORY

R. Drenick
Radio Corporation of America
Camden, N. J.

Introduction

The theory discussed in this paper deals with the problem of the synthesis of certain filters which either extract signals from noise, or extropolate them, in some optimum fashion. Its main purpose is a study of the conditions under which these optimum filters are non-linear, by which procedures they can in principle be synthesized, and how much better they can be expected to perform than their linear counterparts.

A prediction theory, in the sense in which the word is used in the communications field, is usually specified by what assumptions are made in three areas:

1. The nature of the signal.
2. The statistics of the noise
3. The error criterion.

To these, it will be convenient to add here a fourth, the principle of data acquisition. The assumptions which characterize the present theory are explained in some detail in section (Ia) below. They are briefly the following:

1. As regards the signal, it is assumed that it is representable, over the period of time of interest, by a polynomial in time. The order of the polynomial is assumed known before hand, but its coefficients are assumed unknown. This assumption underlies also a prediction theory due to R. B. Blackman, A. W. Bode, and C. E. Shannon (Ref. 1) which is intended for the design of radar tracking filters.

2. The theory is considered more general, for one, as regards to noise. Unlike its predecessor, it is not limited to Gaussian noise but accommodates a very broad class of probability distributions, presumably broad enough to cover most practical applications. It is particularly the cases of non-Gaussian noise which are found to usually lead to non-linear prediction filters.

3. Greater generality can also be claimed with some justification for the assumption concerning the error criterion. Part of the paper, i.e. namely that dealing with the general theory, holds for a rather large variety of error criteria. Another part, however, is restricted to the rms error criterion because explicit results can be derived most readily for this case.

4. As regards the method of data acquisition, the present theory is probably to be considered more restricted than many earlier ones: Unlike that of Blackman, Bode, and Shannon, for instance, which holds for continuous data acquisition, the present one holds (at least formally) only for discontinuous acquisition. The resulting filters are, accordingly, in the nature of sampling filters.

Aside from the reference mentioned, the work reported in this paper has profited from two other sources. One is a particularly elegant treatment of a statistical estimation problem by M.A. Girshick and L.J. Savage (Ref. 3). In fact the general approach, and even the notation used here, borrows heavily from these two authors. The second source is an unpublished memorandum by P. Nesbeda and the Author (Ref. 2) in which very similar though somewhat less general results were arrived at by a different method. This study was very helpful to the one reported here.

One other fact may be worth mentioning. The present theory is a part of statistical estimation theory and as such is not too closely related to N. Wiener's prediction theory (Ref. 5). (The latter, in a similar situation as here, could be obtained as a "Bayes" solution of the prediction problem Ref. 6).

On the basis of the assumptions listed above, the theory is developed in these steps: A functional equation is first introduced which in fact defines what is meant by a predicting (or smoothing) filter. Several preliminary concepts are next introduced which are connected with this prediction concept, and needed in the later discussions. The equation is then introduced which characterizes the predictions which is optimum relative to a given type of signal and a given error criterion. A proof is given of its optimum performance.

The remainder of the paper is concerned with the conveniently simple rms error criterion. Linear filters are established for the case of the Gaussian noise, and a recurrence relation is shown to exist among them. They are also used as starting points for the developments of the non-linear filters which result for non-Gaussian noise. The synthesis method which is obtained for these filters is fairly straightforward and is programmed into a six step procedure. The procedure is illustrated by an especially simple example, namely, that of a linearly varying signal embedded in weakly non-Gaussian noise. The improvement is calculated which is obtained from the non-linear filter over the best linear filter. This improvement seems to be quite substantial.

Ia Assumptions

As pointed out in the Introduction, it is necessary to specify the nature of the prediction problem by assumptions in four areas, namely, the type of signal, the character of the noise, the error criterion, and the manner of data acquisition. It will be convenient to discuss these assumptions now and, at the same time, to define some terminology which will be used in the remainder of the paper.

The pure signal, first of all, will be assumed to be in the form of a polynomial in time.

$$\bar{x}(t) = \theta_o + \theta_1 t + \ldots\ldots + \theta_q t^q. \qquad (o \leq t \leq \bar{t}) \qquad (1.1)$$

This is to say it is assumed that a record of the signal in the absence of noise, taken over a period of at most $\bar{t}$ seconds, can always be fitted with a q -order polynomial, with an error which is negligible for the purpose under study. This is the assumption which underlies the theory of Blackman, Bode, and Shannon (Ref. 1).

It will become evident below that the design of the predicting filter depends very strongly on what is assumed for the order of the polynomial (1.1). Accordingly, it will be convenient to speak of a "q - order" predicting filter, or more briefly a "q - order predictor" when the order of the polynomial (1.1) is q.

The noise, to state the second group of assumptions, will be assumed additive. That is to say, the actually observed signal $x(t)$ contaminated with noise $\epsilon(t)$ will be of the form

$$x(t) = \bar{x}(t) + \epsilon(t)$$

The noise will be characterized by a multivariate probability density, say,

$$f(\epsilon_o, \epsilon_1, \ldots\ldots \epsilon_n) = f(x_o - \bar{x}_o, x_1 - x_1, \ldots\ldots x_n - \bar{x}_n), \qquad (1.2)$$

describing the joint distribution of $(n+1)$ noise samples taken at $(n+1)$ different times. Concerning f, it will be assumed that it can be expanded into a multivariate Gram-Charlier series of type A. This constitutes a restriction on the generality of f which will be rarely important in practice. (It is, in fact, unnecessarily narrow for most of the topics of this paper, but is useful in its last part).

The next, and third set of assumptions deals with the error criterion. It is customary in this connection to prescribe the so-called rms error criterion. This means the following: A penalty is, in effect, introduced for an error in each prediction which varies as the square of that error. The mean value of this penalty is designated as the "mean square error" and the performance of various predicting filters when rated under this rms error criterion, is rated according to this mean square error. The optimum filter is, accordingly, one which minimizes it.

The generalization of this criterion suggests itself, and is, in fact, quite well known (Ref. 6). For the purpose of this paper, the penalty W will be assumed to be a function only of the prediction error ϵ_p

$$W = W(\epsilon_p) \qquad (1.3)$$

The error criterion, that is the basis on which to rate predicting filters, will then be the mean value of W. The penalty W is traditionally called the "loss function". Its mean value is called the "risk" and will be denoted.

$$r = E\{W(\epsilon_p)\} \qquad (1.4)$$

where the symbol E, as customary, stands for "the mean value of". In this nomenclature then, the criterion of performance of a predictor will be the risk to which it leads. The optimum predictor will be the one which minimizes the risk.

In this paper, the loss function will be subject to the natural assumption of being the smallest when the prediction error is zero. In addition to that it will be assumed convex and twice differentiable at the origin. Thus, it includes as a special case the squared error loss. In fact, one derives directly from these assumptions that

$$W'(\epsilon_p) = \epsilon_p V(\epsilon_p) \tag{1.5}$$

where

$$V'(0) \geq 0 \tag{1.6}$$

The rms error criterion is, therefore, characterized by

$$W(\epsilon_p) = \epsilon_p^2, \quad V(\epsilon_p) = 2 \tag{1.7}$$

It will be convenient in what follows to disregard the possibility of the equality sign in (1.5). This simplifies the statements which will be made without materially affecting their principle.

The fourth and final set of assumptions concerns the method of data acquisition. The usual assumption, and also the one leading to the most elegant results, is that of continuous acquisition. Unfortunately, this does not seem possible in the present case, at least if involvement in fairly complicated non-linear functionals is to be avoided. It will, accordingly, be assumed here that data are acquired in a discrete sequence. As a matter of convenience (but not of necessity), a second one will be introduced, namely, that acquisition takes place at a uniform rate. That is to say, the time interval T between any two data points is the same. However, it is not important how short this interval is. The rate of acquisition can, in principle, be taken as high as may seem desirable.

Ib. Terminology

Two time intervals have been introduced above: $\bar{t}$ the period over which (1.1) is an adequate representation of the signal, and T, the time interval between data points. Hence, if

$$\bar{t} \doteq nT \qquad (n = integer)$$

there will be $(n+1)$ data points available on which to make the predictions $x_0, x_1, \ldots x_n$. We shall assume that x_0 denotes the most recent of these points, and that the sequence has been recorded, respectively, at the times

$$t_j = -jT \qquad (j = 1, \ldots\ldots n)$$

(This notation implies, that the origin of time coincides with the most recent observation). We reserve the symbols j and K, as far as possible, for an integer which assumes, as above, the values from 0 to n.

Now, the pure signal is given by (1.1) If it had been sampled at the times t_j the values

$$\bar{x}_j = \theta_o + \theta_1(-jT) + \theta_2(-jT)^2 \ldots\ldots + \theta_q(-jT)^q$$

would have been obtained. The predicted value of the signal, that is, the value of $\bar{x}$ at the time t_p, would be, by (1.1.),

$$x_p = \theta_o + \theta_1 t_p + \theta_2 t_p^2 + \ldots + \theta_q t_p^q$$

If t_p is set equal to zero, the prediction problem reduces to the smoothing problem. One can, therefore, establish as the relation between the desired prediction and the original values observed in the absence of noise

$$\bar{x}_j = x_p + \theta_1[-jT - t_p] + \theta_2[(-jT)^2 - t_p^2] + \ldots + \theta_q[(-jT)^q - t_p^q] \tag{1.8}$$

It will be convenient to put

$$(-jT)^i - t_p^i = a_{ij} \tag{1.9}$$

and write for (1.8)

$$\bar{x}_j = x_p + \sum_{i=1}^{q} a_{ij}\,\theta_i \tag{1.10}$$

The symbol i, unless specified otherwise, will be used consistenly in this paper to denote an integer which takes the values from 1 to q.

One further convention will be useful. We shall frequrntly have to deal with functions of all $(n+1)$ values of x, their joint probability density being one of them. It will be convenient to write, for instance,

$$f(\epsilon_1,\ \epsilon_2,\ \epsilon_n) = f(\epsilon_j)_o^n \tag{1.11}$$

In this notation, and using (1.10), one can re-write (1.2):

$$f(\epsilon_j)_n^o = f\,(x_j - \bar{x}_j)_o^n = f(x_j - x_p - \sum_i a_{ij}\,\theta_i)_o^n \tag{1.12}$$

The same notation will be convenient also in other funtions of these $(n+1)$ variables.

IIa Definition and Characterization of Predictors

The term predictor, or predicting filter, as used in this paper, will be a device, subject to ertain restrictions, which accepts signal and noise as input, stores this input, and generates as ts output an estimate of the signal value at some future time. It will generate this output from .he input data x_j, past and present, according to some formula which will here be called the pre-lictor formula or, when confusion with the physical device is unimportant, briefly the predictor. The predictor formula, then, will be in the nature of

$$u = u(x_o,\ x_1,\ \ldots\ .\ x_n) = u\,(x_j)_o^n$$

(Using again the notation of (1.11).

As mentioned earlier, this formula will depend on the order q of the polynomial which repres-ents the signal. This fact will be expressed by a subscript q. That is to say,

$$u_q = u_q\,(x_j)_o^n$$

stands for a predictor of order q. A second order predictor u_2 according to this terminology, will be designed to extrapolate signals which vary, at the most, quadratically with time. It will, of course, do as well with signals which vary linearly with time or remain constant since these are only special cases of the signal for which it is designed. But it will be inadequate, for instance, with any signal which varies as the cube of time, in the sense that the prediction error will not be zero, even in the absense of noise $(\epsilon_p = u_q - x_p)$.

These statements must now be made more precise. It is necessary, in other words, to specify what is meant by a predictor being "designed", or "adequate", for a q-order signal.

To do this, we proceed with the following argument: Assume that a set of data points x_j has been observed, that they have been put into a predictor of order q and that the output n_q has been obtained. Assume next that a second set x_j' is formed from the first by

$$x_j' = x_j + x_p + \sum_i a_{ij}\,\theta_i \tag{2.1}$$

that is, by superposing on the first set another polynominal of order q. We shall require (as seems natural) that, in this case, the predictor output should be correspondingly increased by x_p:

$$u_q(x_j')_o^n = u_q\,(x_j + x_p + \sum_i aij\ \theta_i)_o^n = u_q\,(x_j)_o^n + x_p \tag{2.2}$$

A filter satisfying this functional equation will be said to have the "predictor property," and it will be prescribed for all predictors discussed in this paper. Equ. (2.2), while common to all predictors (of the same order q) does not determine some specific predictor formula. For, there are many formulae which satisfy it, and some will be suggested presently.

Requirements similar to (2.2), are found in all smoothing and prediction theories, although the reasons by which they are justified are not always the same. The one used here is parallel to the argument first used apparently by Pitman (Ref. 9). Zadeh and Ragazzini (Ref. 6) derive it from the specification that the means of input and output should be the same (a specification which cannot always be used for error criteria other than the least - squares criterion). Schoenberg (e.g., Ref. 10) uses the term "preserving power" for the condition analogous to (2.2).

The predictor property has an immediate consequence which is virtually equivalent to it. To derive it, assume that the originally recorded set of data points x_j has yielded all zeros. Then

$$u_q(x_j)_o^n = u_q(\bar{x}_j)_o^n = u_q(x_p + \sum_i a_{ij}\theta_i)_o^n = u_q(o)_o^n + x_p$$

For those predictors, therefore, which produce no output for zero input (and most will be of that type), the property (2.2) prescribes that they should extrapolate exactly when no noise is present.

It may be worth noting also that the predictor property can be given another interpretation with a rather mathematical flavor: One can, first of all, think of (2.1) as a linear transformation. Equ. (2.2) then expresses a certain invariance property under that transformation which all predictors must have. This is a convenient, if rather abstract, way of expressing the predictor property and will be used repeatedly in what follows.

Finally, we state without proof a few facts which are chiefly of statistical interest. It can be shown that the risk r from any predictor having the property (2.2) is constant, that is, independent of x_p and θ_i. In fact, r can be obtained formally by calculating the mean loss for vanishing x_p and θ_i. This will be expressed (following the notation of Girshick and Savage) by writing E_o for the mean in such a case instead of E as in (1.4)

$$r = E\{W(\epsilon_p)\} = E_o\{W(u_q)\} \tag{2.3}$$

An optimum predictor, namely, the one that minimizes the risk, minimizes this constant. By some well known theorems (e.g., Ref. 6), this makes the optimum predictor the minimax solution to the prediction problem.

IIb Linear Predictors

It may be useful to illustrate the above said by two examples which are presented below. It will be noticed that both are linear predictors in the sense that the predictor formulae are linear in the x_j. Formulae like these can be, and have been, derived for the case of continuous data acquisition by the method of Ref. 1. Set 2, is particular, is the discontinuous analog of optimum filters obtained by that method for the case of white Gaussian noise.

Examples of predictor formulae are:

SET 1:

$$u_o(x_j)_o^n = x_o$$

$$u_1(x_j)_o^n = ax_o + \beta x_1, \quad a = 1 - \frac{t_p}{T}, \quad \beta = \frac{t_p}{T}$$

$$u_2(x_j)_o^n = ax_o + \beta x_1 + \gamma x_2, \quad a = 1 + \frac{3}{2}\frac{t_p}{T}\left(1 - \frac{t_p}{T}\right), \quad \beta = \frac{t_p^2}{T^2} - 2\frac{t_p}{T}, \quad \gamma = -\frac{1}{2}\left[\frac{t_p^2}{T^2} - \frac{t_p}{T}\right]$$

etc.

SET 2:

$$u_o(x_j)_o^n = \frac{1}{n+1}\sum_{j=o}^{n} x_j$$

$$u_1(x_j)_o^n = \frac{\begin{vmatrix} \Sigma x_j & -\Sigma a_{1j} \\ -\Sigma a_{1j}x_j & \Sigma a_{1j}^2 \end{vmatrix}}{\begin{vmatrix} n+1 & -\Sigma a_{1j} \\ -\Sigma a_{1j} & \Sigma a_{1j}^2 \end{vmatrix}}$$

Set 2 (continued)

$$u_2(x_j)_o^n = \frac{\begin{vmatrix} \Sigma x_j & -\Sigma a_{1j} & -\Sigma a_{2j} \\ -\Sigma a_{1j}x_j & \Sigma a_{1j}^2 & \Sigma a_{1j}a_{2j} \\ -\Sigma a_{2j}x_j & \Sigma a_{1j}a_{2j} & \Sigma a_{2j}^2 \end{vmatrix}}{\begin{vmatrix} n+1 & -\Sigma a_{1j} & -\Sigma a_{2j} \\ -\Sigma a_{1j} & \Sigma a_{1j}^2 & \Sigma a_{1j}a_{2j} \\ -\Sigma a_{2j} & \Sigma a_{1j}a_{2j} & \Sigma a_{2j}^2 \end{vmatrix}}$$

etc.

It is easy to convince oneself that these sets actually have the predictor property. A method by which these, and similar ones, can be derived will become clear later in this memo (Section III). The examples cited above are probably sufficient to facilitate their formal extension to higher orders than the second.

Linear predictors such as these form a fairly large class which will play an important role in the theory below. They will, in fact, be so often used as to make a special notation for them advisable. In this paper, a linear predictor of order q will be denoted by v_q, an arbitrary predictor, by u_q.

One further item: It will be important to derive an expression for the output of a q-order predictor, particulary a linear one, where its input is a polynomial of an order r higher than q, and no noise. Assume, more specifically, that a given predictor is linear and of order q, v_q, and that the input is of the form

$$x_j = x_p + \sum_{i=1}^{r} a_{ij}\theta_i \qquad (r > q)$$

Because of its linearity, the predictor will produce an output which will be linearly superposed of several portions. First, there will be the response to input portion

$$x_p + \sum_{i=1}^{q} a_{ij}\theta_i$$

which, due to the predictor property of v_q, will be

$$v_q\left(x_p + \sum_{i=1}^{q} a_{ij}\theta_i\right)_o^n = v_q(0) + x_p$$

Secondly, there will be a group of terms in the output resulting from the input portion

$$\sum_{i=q+1}^{r} a_{ij}\theta_i$$

These will be the form

$$v_q\left[\sum_{i=q+1}^{r} a_{ij}\theta_i\right]_o^n = \sum_{i=q+1}^{r} \theta_i v_q(a_{ij})_o^n,$$

again because of the linearity of the device. The output of a q order predictor with an input of order r greater than q will, accordingly, be

$$v_q\left(x_p + \sum_{i=1}^{r} a_{ij}\theta_i\right)_o^n = v_q(0)_o^n + x_p + \sum_{i=q+1}^{r} \theta_i v_q(a_{ij})_o^n$$

or, slightly more generally,

$$v_q\left[x_j + x_p + \sum_{i=1}^{r} a_{ij}\theta_i\right]_o^n = v_q(x_j)_o^n + x_p + \sum_{i=q+1}^{r} \theta_i v_q(a_{ij})_o^n \qquad (2.4)$$

This relation will be used presently.

IIb. The Auxiliary Quantities z_j

The optimum predictors which are about to be developed will be seen to involve data points x_j only in certain combinations, with quite specific properties. These combinations will be denoted with z_j, in accordance with the notation introduced by Girshik and Savage (Ref. 3).

The quantities z_j can be constructed promptly if one already knows a set of predictors of all orders up to and including the order q of interest. Let it, therefore, be assumed that such a set is known and, more specifically, that it is linear. This set will be denoted with $v_0, v_1, \ldots v_q$. It could, for instance, be one of the two examples listed above in Section IIa. The quantities z_j are then formed as linear combinations of x_j and these v_i; thus,

$$z_j = x_j + \sum_{i=o}^{q} \gamma_{ij} v_i \tag{2.5}$$

The coefficients γ_{ij} are, however, not arbitrary. They are, more specifically, so determined that z_j remain invariant under the transformation (2.1). The procedure by which the γ_{ij} can be determined is straightforward and will be carried out, as a matter of convenience, for the case $q = 2$. The extension to higher orders will be quite obvious from that.

The coefficients γ_{ij} in (2.5) can be determined directly from the requirement of invariance under the above transformation, with $q = 2$. For one must have

$$z_j = x_j + x_p + \sum_{i=1}^{q} a_{ij}\theta_i + \Big\{\gamma_{oj} v_o (x_j + x_p + \sum_{i=1}^{2} a_{ij}\theta_i)_o^n + \gamma_{1j} v_1 (x_j + x_p + \sum_{i=1}^{2} a_{ij}\theta_i)_o^n + \gamma_{2j} v_2 (x_j + x_p + \sum_{i=1}^{2} a_{ij}\theta_i)_o^n \Big\}$$

Using the predictor property of v_0, v_1, v_2, and observing (2.4), this leads to the following system of $(3n+3)$ simultaneous equations for the γ_{ij}

$$\begin{aligned} &\gamma_{oj} + \gamma_{1j} + \gamma_{2j} = -1 \\ &\gamma_{oj} v_o (a_{1j})_o^n = -a_{1j} \\ &\gamma_{oj} v_o (a_{2j})_o^n + \gamma_{1j} v_1 (a_{1j})_o^n = -a_{2j} \end{aligned} \tag{2.6}$$

This derivation of the coefficients γ_{ij} for the case $q = 2$ is clearly and easily extended to higher orders.

In the discussions which follow, it will sometimes be convenient to rearrange the z_i into a different form. If one takes into account the specific solutions obtained for the γ_{ij} from (2.6), one finds promptly that one can write,

For $q = 0$: $z_j = x_j - v_o$

For $q = 1$: $z_j = (x_j - v_1) - \dfrac{a_{1j}}{v_o(a_{1j})}[v_o - v_1]$ (2.7)

For $q = 2$: $z_j = (x_j - v_2) - \dfrac{a_{1j}}{v_o(a_{1j})}[(v_o - v_2 - C_1(v_1 - v_2)] - \dfrac{a_{2j}}{v_1(a_{2j})}[v_1 - v_2]$, where $C_1 = v_o(a_{2j})/v_1(a_{2j})$

For $q = 3$: $z_j = (x_j - v_3) - \dfrac{a_{1j}}{v_o(a_{1j})}[(v_o - v_3) - C_1(v_1 - v_3) - C_2(v_2 - v_3)] - \dfrac{a_{2j}}{v_1(a_{2j})}[(v_1 - v_3) - C_3(v_2 - v_3)] - \dfrac{a_{3j}}{v_2(a_{3j})}[v_2 - v_3]$

where

$$C_2 = \begin{vmatrix} v_o(a_{j3}) & v_1(a_{j3}) \\ v_o(a_{j2}) & v_1(a_{j2}) \end{vmatrix} \Big/ v_1(a_{j2})v_2(a_{j3}),\quad C_3 = v_1(a_{j3}) \Big/ v_2(a_{j3})$$

etc.

The predictors v_o, v_1, v_2, v_3 in these formulae stand, of course, for $v_o(x_j)_o^n, v_1(x_j)_o^n, v_2(x_j)_o^n, v_3(x_j)_o^n$.

IIc. The Optimum Predictor

In this section, the equation will be introduced which determines the optimum predictor of order q for a given loss function. This is a predictor u_q^* which minimizes the risk (2.4) (asterisks will henceforth be used to denote optimum predictors):

$$r = E_o\{W(u_q^*)\} = \min E_o\{W(u_q)\}$$

It may be worth repeating that the loss functions which are admitted here, following the discussion in Section Ia, are of the form

$$W(y) = y\,V(y), \qquad W^{\mu}(y) \geqq 0$$

The equation which determines the optimum predictor will be seen to be an implicit one. It is assumed in establishing it that some complete set of linear predictors, up to and including one of the order q, $v_o, v_1, \cdots v_q$, is already known. Either of the two sets in Section IIb, for instance, could be used (but would not usually be the most appropriate). The optimum predictor u_q^* can then be derived formally from the known one by adding a correction term to the q-order linear predictor v_q:

$$\boxed{u_q^* = v_q + \Delta u} \tag{2.8}$$

The equation under discussion is an implicit equation for this correction term Δu. In fact, it will be shown presently that Δu is determined by

$$\boxed{\Delta u = -\frac{E_o[v_q\,V(v_q + \Delta u)\,|z_j]}{E_o[V(v_q + \Delta u\,|\,z_j)]}} \tag{2.9}$$

Here, $E_o[V(v_q + \Delta u)\,|z_j]$ stands for the conditional expectation of $V(v_q + \Delta u)$ for given z_j. The z_j are the quantities introduced in the preceding section, and formed using the known $v_o, v_1 \cdots v_q$. The symbol E_o has been introduced in Section IIa as the mean value of a function for $x_p = \theta_1 = \theta_2 \ldots = \theta_q = 0$.

Equations (2.8), (2.9) are the main results of the present paper. A corresponding equation for a squared-error loss function, and for what in the present terminology would be called a zero-order predictor, is contained in the paper by Girshick and Savage (Ref. 3).

It must now be proven that equation (2.9) does indeed yield a correction term which turns u_q into the optimum. This proof will proceed in three steps. It will first be shown that u_q^* is again a q-order predictor. The second step will establish an auxiliary fact, namely, that

$$E_o\{W'(u_q^*)\} = 0 \tag{2.10}$$

The third step will finally prove that u_q^* is the optimum, that is, that it minimizes the risk r.

The first step is very direct. Δu, by equation (2.9), is clearly a function of only the z_j which have been constructed to be invariant under the transformation. v_q has the predictor property under the same transformation. Hence, u_q^* has the predictor property.

To establish (2.10), note that, by (1.5) and (1.6)

$$E_o\{W'(u_q^*)\} = E_o E_o\{W'(v_q + \Delta u)\,|z_j\} = E_o E_o\{(v_q + \Delta u)\,V(v_q + \Delta u)\,|\,z_j\} = E_o\{E_o\{v_q\,V(v_q + \Delta u)\,|\,z_j\} + \Delta u\,E_o\{V(v_q + \Delta u)\,|\,z_j\}\}$$

In these expressions, the symbol $E_o E_o$ indicates the mean value is first taken conditionally, for given z_j, and then over the z_j. The term contained in the exterior braces of the last expression is zero, by (2.9). This establishes (2.10).

The last step in the process is the proof of the optimum character of u_q^*. This proof is carried out by showing that the risk involved in using any predictor u_q other than u_q^* cannot be smaller than the risk due to u_q^*. This is seen in the following way:

$$
\begin{aligned}
r(u_q) &= E_o\{W(u_q)\} \\
&= E_o\{W(u_q^*) + (u_q - u_q^*)\, W'(u_q^*) + (u_q - u_q^*)^2\, \theta\, W''(u_q^*)\} \\
&= r(u_q^*) + E_o\{(u_q - u_q^*)\, W'(u_q^*)\} + \tfrac{1}{2} E_o\{(u_q - u_q^*)^2\, W''(\theta u_q)\} \\
&= r(u_q^*) + E_o E_o\{(u_q - u_q^*)\, W'(u_q^*) \mid z_j\} + \tfrac{1}{2} E_o E_o\{(u_q - u_q^*)^2\, W''(\theta u_q^*) \mid z_j\}
\end{aligned}
$$

Now, $u_q - u_q^*$ depends only on z_j because it is invariant under the transformation (2.1). Therefore,

$$
r(u_q) = r(u_q^*) + E_o\{(u_q - u_q^*)\, E_o[W'(u_q^*) \mid z_j]\} + \tfrac{1}{2} E_o\{(u_q - u_q^*)^2\, E_o\{W''(\theta u_q^*) \mid z_j\}
$$

The second of the right-hand terms vanishes because of (2.10), and the third is non-negative because of the convexity of $W(y)$. Hence, the risk $r(u_q)$ with the arbitrary predictor u_q

$$
r(u_q) = r(u_q^*) + \tfrac{1}{2} E_o\{(u_q - u_q^*)^2\, E_o\{W''(\theta u_q^*) \mid z_j\}\} \tag{2.11}
$$

is certainly no smaller than that obtained with u_q^*, and equality prevails only in the unusual cases in which

$$
E_o\{(u_q - u_q^*)^2\} = 0
$$

Thus, u_q^* is indeed an optimum predictor.

Equation (2.9) determining this optimum predictor will often be difficult to mechanize. The cases, for instance, in which $W(y)$ is expressible by a polynomial of higher order than the second, lead to algebraic equations for Δu. This is fortunately, not true for the case of least-square prediction. In this case the equation for Δu is linear and it can be solved explicitly. The remainder of this paper will, accordingly, deal with that specific case.

III. Least-Square Error Prediction

IIIa. Specialization To The RMS Error Criterion

The specialization of the preceding theory to the rms can be achieved readily. In fact, since $V(y) = 1$ in this case, equation (2.9) reduces to

$$
u_q^* = v_q - E_o(v_q \mid z_j) \tag{3.1}
$$

This shows that, in the rms-error case, the optimum predictor is determined by an explicit equation.

It will be useful later on to have the equivalent to equation (2.11) which established u_q^* as the optimum. It is

$$
E_o\{u_q^{*2}\} = E_o\{u_q^2\} = E_o\{(u_q^* - u_q)^2\} \tag{3.2}
$$

It has already become apparent in the theory so far that linear predictors play an important role. This, and certain other features, recommend them for an early discussion. The next section is, in fact, devoted to it.

IIIb. Optimum Linear Predictors

It has been known for some time that the optimum predictor is linear if the probability density of the noise is Gaussian and the loss function is the squared error (Ref. 8). A proof that the same set of linear predictors is also optimum for a large class of symmetric loss functions was recently given by P. Nesbeda and the writer (Ref. 7). Optimum linear predictors can be derived rather expeditiously by the present theory. It is found, more specifically, that one can use a recursion formula to proceed from zero order to higher order linear predictors. This, and their usefulness in the present theory, suggests their derivation here.

Again, as a matter of convenience, the derivation will be developed for limited order q, that is, for $q = o$ and $q = 1$. The extension to predictors of other orders will be quite plain.

By (3.1), any optimum predictor

$$E_o (u_q^\circ | z_j) = O$$

In the case under consideration, this condition can be written (using again $v's$ to denote linear, and asterisks to denote optimum, predictors)

$$\iint \ldots \int v_q^* \, dv_o^* \, dv_1^* \ldots dv_q^* \, \phi (z_j - \sum_{i=o}^{q} \gamma_{ij} v_i^*)_o^n = O \tag{3.3}$$

where $\phi (y_j)_o^n$ stands for the $(n + 1)$ dimensional Gaussian

$$\phi (y_j)_o^n = (2\pi)^{-\frac{n}{2}} \Lambda^{-\frac{1}{2}} \exp \left(-\frac{1}{2\Lambda} \sum_{j,K=O}^{n} \Lambda_{jK} y_j y_K \right)$$

Assume first the order $q = o$. Equation (3.3) can then be rewritten in the form

$$O = \int v_o^* \, \phi (z_j + v_o^*)_o^n \, dv_o^* = \exp \left\{ -\frac{1}{2\Lambda} \Sigma\Sigma \, \Lambda_{jK} z_j z_K \right\} \cdot \int u_o^* \exp \left\{ -\frac{1}{2\Lambda} [u_o^{*2} \Sigma\Sigma \Lambda_{jK} - 2u_o^* \Sigma\Sigma \Lambda_{jK} z_j] \right\} du_o^* \tag{3.4}$$

It is easy to convince oneself, by any number of equivalent arguments, that (3.4) can hold only if

$$\Sigma\Sigma \, \Lambda_{jK} z_j = O \tag{3.5}$$

This relation will be useful in this form later on (Section III). The optimum linear predictor of order zero is, in fact, a direct consequence of (3.5). By putting in it, from (2.5),

$$z_j = x_j - v_o^*$$

one has (3.6)

$$v_o^* = \Sigma\Sigma \, \Lambda_{jK} x_j \Big/ \Sigma\Sigma \, \Lambda_{jK}$$

for the optimum linear predictor of order zero.

The procedure is entirely parallel for the corresponding first-order predictor. The integral in (3.3) is then best written in the form,

$$O = \iint v_1^* \, dv_o^* \, dv_1^* \, \phi [z_j - \frac{a_{1j}}{u_o^* (1)} (v_o^* - v_1^*) - v_1^*]_o^n = \iint v_1^* \, dw_o \, dv_1^* \, \phi [z_j - \frac{a_{1j}}{v_o^* (1)} w_o - v_1^*]_o^n$$

that is, using (2.7). If the same procedure as above is applied to the present integral, one obtains two equations which are necessary for the vanishing of this integral. One of these is again (3.5). The other, namely,

$$\Sigma\Sigma \, \Lambda_{jK} a_{1K} z_j = O \tag{3.7}$$

is the relation among z_j which, for first-order predictors, accompanies (3.5). The predictor formula itself is promptly derived from this by replacing z_j with x_j, using for instance, equation (2.5). One obtains

$$v_1^* \Sigma\Sigma \Lambda_{jK} \lambda_{1j} a_{1K} = \Sigma\Sigma \Lambda_{jK} x_{iK} x_j - v_o^* \Sigma\Sigma \Lambda_{jK} \gamma_{oj} a_{1K} \tag{3.8}$$

as a recursion formula for the optimum linear predictor of first order. The predictor v_o^* could be placed in (3.8) with the expression (3.6).

The extension of this type of predictor formula to predictors of arbitrary order is quite clear. One must have

$$v_q^* \Sigma\Sigma \Lambda_{jK} \gamma_{qj} a_{qK} = \Sigma\Sigma \Lambda_{jK} a_{qK} x_j - v_o^* \Sigma\Sigma \Lambda_{jK} \gamma_{oj} a_{qK} - \cdots - v_{q-1}^* \Sigma\Sigma \Lambda_{jK} \gamma_{(q-1)\,j} a_{qK} \tag{3.9}$$

This is the general recursion formula for an optimum linear predictor of arbitrary order, expressed in terms of those of lower orders.

These expressions become especially simple in those cases in which successive observations are statistically independent, and the noise is stationary. Then

$$\Lambda_{jK} = \delta_{jK} \sigma^2$$

The resulting predictor formulae are then those listed earlier as sample set 2 in Section IIa.

IIIc. Optimum Non-Linear Predictors

When the probability distribution $f(x_j - x_p - \sum_{i=1}^{q} a_{ij}\theta_i)_o^n$ of the noise is not Gaussian the optimum rms predictors are found to be in general non-linear. This will be shown in the present section. It will develop, more particularly, that these predictors can be obtained by a fairly simple procedure in many cases which are likely to be of practical interest. These cases are characterized by the fact that the noise distributions do not differ too radically from the Gaussian. Under such conditions, one can express the actual probability density of the noise as an n-dimensional Gram-Charlier series (of type A), and have reason to hope that only a few terms of this series will be necessary for an adequate representation.

An n-dimensional Gram-Charlier series is one in which the given probability density of the noise is developed into a series starting with a well chosen Gaussian, $\phi(x_j - x_p - \sum a_{ij}\theta_i)_o^n$, and continuing with the derivatives of that Gaussian with respect to all of its variables $x_o, x_1, \ldots x_n$. A "well chosen" Gaussian is one whose first- and second-order moments (i.e., means, dispersions and correlation) agree with those of the given distribution $f(x_j - x_p - \sum a_{ij}\theta_i)_o^n$. An arbitrary term from such a series could be of the form

$$\frac{a(\nu_1, \nu_2, \cdots \nu_n)}{\nu_1!\,\nu_1!\cdots\nu_n!} \cdot \frac{\partial^{\nu_1+\nu_2+\cdots+\nu_n}}{\partial x_1^{\nu_1}\,\partial x_2^{\nu_2}\,::\,\partial x_n^{\nu_n}} \phi(x_j - x_p - \sum_{i=1}^{q} a_{ij}\theta_i)_o^n \tag{3.10}$$

where $a(\nu_1, \nu_2, \ldots \nu_n)$ is a constant. If the Gaussian is in fact chosen as recommended, no derivative of an order lower than the third will appear in the series, and if the actual noise distribution is symmetric, the fourth-order derivatives will be the lowest (other than the Gaussian ϕ itself, of course, which is the zeroth derivative).

For brevity, the notation

$$\frac{a_\nu}{\nu!} \frac{\partial^\nu}{\partial x_j^\nu} \phi(x_j - x_p - \sum_{i=1}^{q} a_{ij}\theta_i)_o^n \tag{3.11}$$

will be used for terms as the one shown above. The Gram-Charlier series for the general noise distribution can then be written symbolically

$$f(x_j - x_p - \Sigma a_{ij}\theta_i)_o^n = \Sigma_\nu \frac{a_\nu}{\nu!} \frac{\partial^\nu}{\partial x_j^\nu} \phi(x_j - x_p - \Sigma a_{ij}\theta_i)_o^n \tag{3.12}$$

The object of the present discussion, is to derive from equation (3.1) optimum q-order predictors for distributions of this type. To do this, as mentioned before, one must have a complete set of linear predictors of all orders, from zero up to and including q. These predictors can be arbitrarily chosen; for instance, either of the sets in Section IIa would serve for the purpose. The necessary procedure is, however, the most expeditious if this starting set of predictors is well chosen. Particularly well suited is, in this case, the "corresponding" set of optimum linear predictors. That is, the set $v_0, v_1 \ldots v_q$ which would be optima if in the Gram-Charlier series (3.12) all terms but the first were zero $(a_\nu = o, \nu > o)$.

Let it be assumed in what follows that such a set has been determined as, indeed, it always could be by the procedure outlined in the preceding section. The remaining task is then the calculation of the conditional expectation in (3.1).

$$E_o(v_q \mid z_j) = \frac{\iint \cdots \int v_q \, dv_1 \, dv_2 \cdots dv_q \, f(z_j - \sum_{i=o}^{q} \nu_{ij} \nu_i)_o^n}{\iint \cdots \int dv_1 \, dv_2 \cdots dv_q \, f(z_j - \sum_{i=0}^{q} \nu_{ij} \nu_i)_o^n} = \frac{N}{D} \tag{3.13}$$

In this equation, the asterisks are not shown for the v's to indicate that they are not optima for the distribution (3.12).

In order to simplify the notation in what follows, the proof will be carried out for the special case $q = 1$. As before, its extension to arbitrary q will be quite evident.

Equation (3.13) shows that the optimum predictor will usually involve a fraction N/D, and it is necessary to establish a procedure by which to calculate it. This procedure will be seen to be quite simple, particularly so for the denominator D. The integral in D can, using the series notation (3.12), be written

$$D = \sum_{\nu} \frac{a_\nu}{\nu!} \iint dv_o \, dv_1 \frac{\partial^\nu}{\partial z_j^\nu} \phi \left[z_j - \frac{a_{ij}}{v_o(1)} (v_o - v_1) - v_1 \right]_o^n$$

Now, v_o and v_1 are predictors which would be optimum for the linear problem. This can be made use of, and the denominator transformed into

$$D = K \sum_{\nu} \frac{d_\nu}{\nu!} \frac{\partial^\nu}{\partial z_j^\nu} \phi (z_j)_o^n$$

where K is a proportionality constant of no relevance to the present problem.

The rule of formation of D in the predictor formula is, accordingly, very simple and patently true regardless of q. In the series expansion of the noise distribution $f(\epsilon_j)_o^n$, replace ϵ_j by z_j. The result is,

$$D = f(zj)_o^n \tag{3.14}$$

The rule for the numerator N is only a trifle more complicated. The integral in N is

$$N = \sum_{\nu} \frac{a_\nu}{\nu!} \iint v_1 \, dw_o \, dv_1 \frac{\partial^\nu}{\partial z_j^\nu} \phi \left(z_j - \frac{a_{1j}}{v_o(1)} w_o - v_1 \right)_o^n$$

One finds, by the same argument as for D,

$$N = K \sum_{\nu} \frac{a_\nu}{\nu!} \frac{\partial^{\nu-1}}{\partial z_j^{\nu-1}} \phi (z_j)_o^n$$

The rule of formation of the numerator is, accordingly, the following. In the Gram-Charlier expansion (3.12) of $f(x_j)_o^n$, omit the first term (the Gaussian proper), and in all others reduce by one the order of the derivative with respect to one of the variables. Finally, replace x_j with z_j. This rule is evidently the same regardless of the order q of the desired predictor.

The conclusion is then reached thát the formula for the optimum non-linear predictor of order q is

$$u_q^* = v_q + \frac{\sum \frac{a_\nu}{\nu!} \frac{\partial^{\nu-1}}{\partial z^{\nu-1}} \phi (z_j)_o^n}{f(z_j)_o^n} \tag{3.16}$$

The procedure by which it can be synthesized can now be outlined in the following steps:

a) Develop the given probability density $f(x_j)_o^n$ of the noise into a generalized Gram-Charlier (Type A) series.

b) First omit all terms from it but the first (which is a Gaussian) and derive a set of optimum linear predictors for this Gaussian. For the derivation, use the recursion process which is represented by equation (3.9). Find also the necessary quantities $v(a_{ij})$ introduced in (2.4).

c) Use this set to form the quantities z_j by the procedure suggested in Section IIc, or by equations (2.7)

d) Substitute these z_j in place of the x_j into the expression for the probability density $f(x_j)_o^n$ of the noise. This is the denominator of the fraction in the desired predictor formulae.

e) Next, omit the first term of the Gram-Charlier series, and lower by one the derivatives with respect to one (any one) of the variables in every other term. Substitute z_j for the x_j. This is the numerator of the fraction of the predictor formula.

f) Insert the fraction as the second right-hand-side term into (3.16). For the first term, use the q-order linear predictor obtained under b).

IIId. Illustrative Derivation of a Simple Non-Linear Predictor

The process of establishing a non-linear predictor by this method is fairly straight-forward, as is evident from the outline in the preceding section. It will now be illustrated with a simple example (which was also used in Ref. 2), namely the determination of a certain first-order predictor u_1^*.

Let it be assumed that the noise source is virtually white, or that the data acquisition is relatively slow, so that successive noise samples can be considered independent. The joint probability density of the noise $f(\epsilon_j)_o^n$ (equation (1.2)) can then be written as a product

$$f(\epsilon_j)_o^n = f_1(\epsilon_1)\, f_1(\epsilon_2) \cdots f_1(\epsilon_n)$$

(Here, and in the discussion below, the subscript 1 will be used to point out a uni-variate probability density.) Assume furthermore, that $f_1(\epsilon_j)$ is symmetrical and weakly non-Gaussian. That is

$$f_1(\epsilon_j) = \phi_1(\epsilon_j) + \frac{a_4}{4!} \frac{\partial^4}{\partial \epsilon_j^4} \phi_1(\epsilon_j)$$

where a_4 is so small that its square can be neglected relative to unity. This is not as much of a restriction as it might seem. In practice, it will often be possible to let a_4 vary over a range from -5 to $+5$ before the approximations to be used here become altogether indefensible. The range of shapes of probability densities which can be produced by this variation in a_4 is shown in fig. 1.

For small a_4 then, one can write approximately

$$f(\epsilon_j)_o^n = \left[1 + \frac{a_4}{4!} \sum_j \frac{\partial^4}{\partial \epsilon_j^4} \right] \prod_K \phi_1(\epsilon_K). \tag{3.17}$$

This puts the probability distribution into the desired form of a Gram-Charlier series, and step (a) in the procedure has been affected.

The next step is to establish the optimum linear predictors of zero and first order, v_o and v_1. Since the successive noise samples are statistically independent, the remark at the end of Section IIIb applies. According to this remark, the required optimum linear predictors are given by set 2 of Section IIa, that is,

$$v_o = \frac{1}{n+1} \sum x_K, \qquad v_1 = \frac{\begin{vmatrix} \Sigma x_K & -\Sigma a_{1K} \\ -\Sigma a_{1K}\, x_K & \Sigma a_{1K}^2 \end{vmatrix}}{\begin{vmatrix} n+1 & -\Sigma a_{1K} \\ -\Sigma a_{1K} & \Sigma a_{1K}^2 \end{vmatrix}}$$

Also needed in this connection is the response of the zero-order filter to a linear input which, in the present case, is

$$v_o(a_{1j})_o^n = \frac{1}{n+1} \sum_{j=o}^{n} a_{1j}$$

The third step in the procedure is to determine the quantities z_j. They are, by equations (2.7)

$$z_j = x_j - \frac{a_{1j}}{v_o(a_{1j})_o^n}(v_o - v_1) - v_1 = \sum_{K=o}^{n} \left[\delta_j^K - \frac{a_{1j}}{S_1} \frac{1}{p} 1 - \frac{n}{S_1} a_{1j} \quad 1 - \frac{S_1}{S_2} a_{1K} \right] x_K$$

Here, δ_j^K is the Kronecker delta, and S_1, S_2 and p stand for

$$S_1 = \sum_{K=o}^{n} a_{1K} = -(n+1)\, T \left(-\frac{n}{2} + \frac{t_p}{T} \right)$$

$$S_2 = \sum_{K=o}^{n} a_{1K}^2 = (n+1)\, T^2 \left[\frac{n(2n+1)}{6} + \frac{nt_p}{T} + \frac{t_p^2}{T^2} \right]$$

$$p = n - \frac{S_2}{S_1^2}$$

This completes the preliminaries.

The numerator N and denominator D for the fraction in u_1^* are obtained by steps e) and d), respectively, of the procedure. They are

$$N = \frac{a_4}{4!} \sum_{j=o}^{n} \frac{\partial^3}{\partial z_j^3} \prod_{K=o}^{n} \phi_1(z_K), \qquad D = \left[1 + \frac{a_4}{4!} \sum_{j=o}^{n} \frac{\partial^4}{\partial z_j^4} \right] \prod_{K=o}^{n} \phi_1(z_K)$$

The desired non-linear predictor is, accordingly,

$$u_1^* = v_1 + \frac{\frac{a_4}{4!} \sum_j \frac{\partial^3}{\partial z_j^3} \prod_K \phi_1(z_K)}{\left[1 + \frac{a_4}{4!} \sum_j \frac{\partial^4}{\partial z_j^4} \right] \prod_K \phi_1(z_K)} \doteq v_1 + \frac{a_4}{4!} \sum_j H_3(z_j) \tag{3.20}$$

Here, use has been made of the smallness of a_4 which renders negligible the second term in the denominator. $H_3(z_j)$ is the third-order Hermite polynomial:

$$H_3(z_j) = z_j^3 - 3z_j$$

where z_j is the linear combination (3.18) of the inputs. v_1 is the optimum linear predictor quoted above.

The predictor (3.20) is very similar, though slightly simpler than one obtained earlier (in Ref. 2). It is interesting to note in this connection that u_1^* is not unique. That is to say, there exist several expressions, all of which have the same optimum characteristics, namely, have the same low rms error. Most of these involve the third-order Hermite polynomial, and all of them have non-linear portions depending on z_j only. Their algebraic structures, however, differ somewhat.

In fact, one such variant which is slightly simpler than (3.20) can be derived promptly. It follows from equation (3.5) that the linear term in

$$\sum_{j=o}^{n} H_3(z_j) = \sum_{j=o}^{n} z_j^3 - 3\sum_{j=o}^{n} z_j$$

can be omitted and

$$u_1^* = v_1 + \frac{a_4}{4!} \sum_j z_j^3 \tag{3.21}$$

be written for the optimum predictor. An illustrative mechanization of the corresponding filter is shown in fig. 2.

Evaluation of a Simple Non-Linear Predictor

It is of interest to investigate the performance of a non-linear predictor such as the one given by equation (3.21). That is to say, one might ask what is to be gained in the magnitude of the rms error from using the optimum non-linear predictor u_1^* as compared to an optimum linear one, namely, v_1.

The best way, such as it is, of carrying out this comparison is by equation (3.2). According to it, and to (3.21), the rms errors of the outputs of u_1^* and v_1 are connected by

$$E_o(u_1^*)^2 - E_o(v_1^2) = E_o[(u_1^* - v_1)^2] = \left(\frac{a_4}{4!}\right)^2 E_o\left[\sum_{j=o}^{n} z_j^3\right]^2$$

Hence, to obtain the desired comparison, one must evaluate the integral

$$E_o\left[\left(\sum_{j=o}^{n} z_j^3\right)^2\right] = \iint \cdots \int \left[\sum_j z_j^3\right]^2 \prod_{K=o}^{n} \phi_1(x_K)\, dx_o\, dx_2 \cdots dx_n \tag{3.22}$$

where z_j is given by (3.18). No method has so far been found by which this integration could be carried out neatly and smoothly. This is apparently a rather common difficulty in statistics, and the case under consideration is no exception, despite its simplicity. One can proceed a few steps algebraically, at the price of mounting involvement, but ultimately numerical work must be resorted to. This was done and led to the result shown in fig. 3. The quantity plotted there is, more specifically,

$$\Delta\epsilon/\epsilon = \sqrt{[E_o(u_1^{*2}) - E_o(v_1^{*2})]/E(v_1^2)}$$

It can be interpreted as the improvement in the least-square error which is achieved by using the optimum non-linear predictor rather than the corresponding linear predictor. This improvement is seen to be by no means negligible, even if the memory of the filter is limited to but a few data points.

References

1) R.B. Blackman, A.W. Bode and C.E. Shannon, "Data Smoothing and Prediction in Fire-Control System", Tech. Report of Division 7, NDRC, Vol. 1 (1946).

2) R. Drenick and P. Nesbeda, "A Preliminary Study of Optimum Non-linear Prediction", unpublished memorandum, dated 4-30-53.

3) M.A. Girshick and L.J. Savage, "Bayes and Minimum Estimates for Quadratic Loss Functions", Second Berkeley Symposium on Mathematical Statistics and Probability, Univ. Calif. Press, 1951, p. 53.

4) L. Zadeh, and J.R. Ragazzini, "Extension of Wiener's Theory of Prediction", Jour. of Applied Phys. 21(1950), p.645.

5) N. Wiener, "Extrapolation and Interpolation of Stationary Time Series", John Wiley & Sons, 1949.

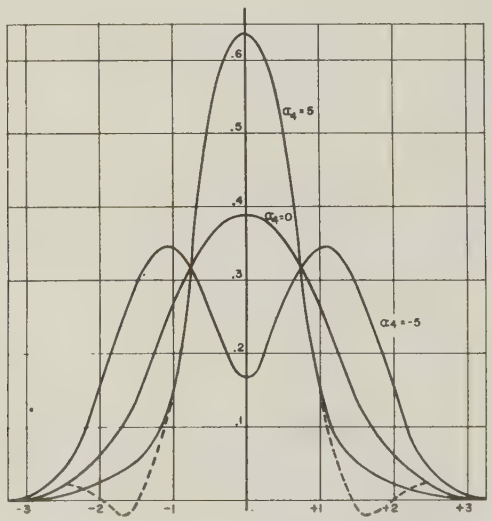
6) A. Wald, "Statistical Decision Functions", John Wiley & Sons, 1950.

7) R. Drenick and P. Nesbeda, "A Class of Optimum Linear Predictors", Paper presented at the Washington meeting of Institute Mathematical Statistics, Apr. 1953.

8) H.E. Singleton, "Theory of Nonlinear Transducers", Tech. Report No. 160, Research Lab. of Electronics, Mass. Inst. Tech., 8-12-50.

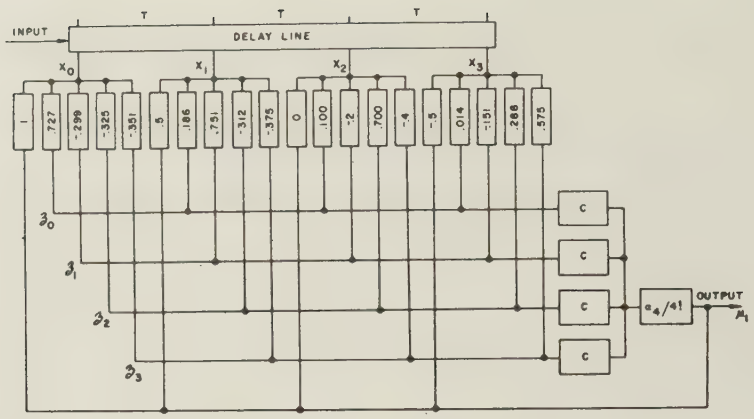
9) E.J.G. Pitman, "The Estimation of Location and Scale Parameters", Biometrika 30 (1939) p. 391.

10) I.J. Schoenberg, On Smoothing Operations and Their Generating Functions, Bull. Am. Math. Soc. Vol. 59 (1953) p.199.



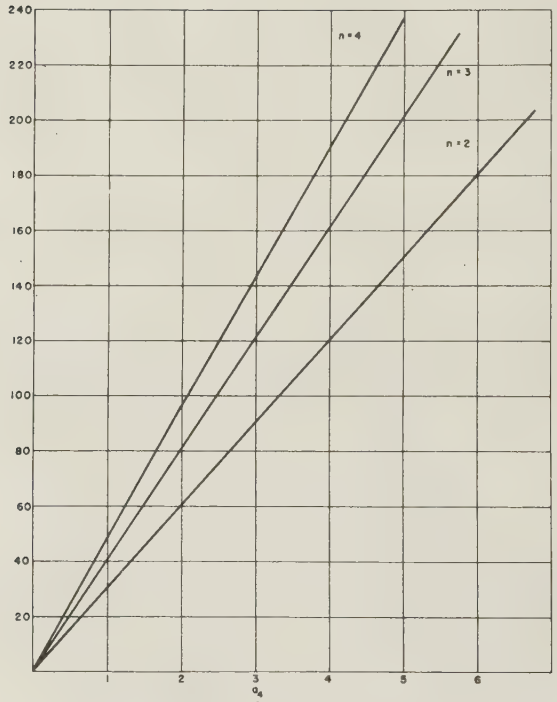
EXAMPLES OF "WEAKLY" NON-GAUSSIAN DISTRIBUTIONS ($\sigma = 1$)

Fig. 1



ILLUSTRATIVE MECHANIZATION 1ST ORDER PREDICTOR, "WEAKLY" NON-GAUSSIAN NOISE, PREDICTION TIME $t_p = T$ (C... CUBING DEVICE)

Fig. 2



PERCENT IMPROVEMENT $\frac{\Delta\epsilon}{\epsilon}$ IN THE RMS ERROR DUE TO TRACKING WITH A NONLINEAR FILTER

Fig. 3

THE DETECTION OF SIGNALS PERTURBED BY SCATTER AND NOISE[α]

Robert Price[β]

Student Member, IRE
Research Laboratory of Electronics
and
Lincoln Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts

Introduction

In recent years, much study has been devoted to the problem of conveying information efficiently through channels in which the message-bearing waveforms may undergo distortion. Statistical methods have proven an effective tool with which to analyze and synthesize transmission systems as a whole, especially where channel perturbations are of a random nature. The statistical approach is particularly rewarding when questions of receiver optimization are considered. Provided that the transmitter and channel conform to the realistic, yet very general, model proposed by Shannon,[1] the ideal receiver assumes the form of a probability-computer. This result was recognized by Woodward and Davies,[2] and Van Vleck and Middleton[3] have similarly treated ideal detection as the testing of statistical hypotheses.

Random channel disturbances which have already received considerable attention have been mostly of an additive nature; shot- and thermal- noise, atmospheric impulse-noise, and adjacent-channel interference are familiar examples in this category. An excellent example of the application of probability-computing methods is given in a paper by Reich and Swerling,[4] who have found the functional form of the ideal receiver for the thermal- or shot- noise case. Recently, studies of VHF "scatter" propagation beyond the line-of-sight have led to a channel model in which disturbances are encountered which are no longer of a purely additive nature. As in the case of additive gaussian noise, it has been found possible to obtain explicitly the functional form of the ideal, probability-computing receiver for the "scatter" channel. The analysis is set forth in the three sections of this paper. In Section A the mechanism by which the channel perturbs a transmitted signal is studied in detail, so that the equivalent mathematical operations may be completely specified. Section B then applies the results of Section A to the exact derivation of the functional form of the probability-computing receiver. Finally, Section C discusses the simplification in receiver design which is made possible by certain approximations valid at small signal-to-noise ratios.

A. Study of the Scatter Channel

For VHF transmissions beyond the line-of-sight, Booker and Gordon[5] have postulated a tropospheric scattering model. Their hypothesis is that the transmitted wave impinges upon a great number of randomly moving irregularities in the troposphere, so that the received wave is the resultant of many small scattered waves. Booker, Ratcliffe and Shinn[6] have proposed a similar model in connection with ionospheric propagation. This paper does not propose to discuss the validity of this model, but will accept it merely as a form of disturbed channel interesting in its own right, regardless of whether it truly exists in nature or not.

α This paper is excerpted from "Statistical Theory Applied to Communication through Multipath Disturbances", an Sc.D. thesis submitted to the Department of Electrical Engineering, Massachusetts Institute of Technology, on 24 August 1953. The research reported in this paper was supported jointly by the Army, Navy, and Air Force under contract with the Massachusetts Institute of Technology.

β Presently at the Commonwealth Scientific and Industrial Research Organization, Sydney, N.S.W., Australia.

The analysis proceeds on the basis of the particular scattering model considered by Rice,[7] in which the randomly-moving irregularities are assumed all of equal size. Taking a sine wave as the transmitted signal, $x_0(t) = \sin \omega_0 t$, it can be shown through Rayleigh's "random walk" analysis[8] and the central limit theorem that the received signal $z_0(t)$ has a gaussian distribution in all dimensions. That is, $z_0(t)$ is statistically identical to filtered thermal noise having the same power spectrum. Assuming that this spectrum $Y(\omega)$ is symmetric and narrow-banded relative to its center frequency ω_0, we may write, from Rice,[9]

$$z_0(t) = y_s(t) \sin \omega_0 t + y_c(t) \cos \omega_0 t \tag{1}$$

$y_s(t)$ and $y_c(t)$ are independent gaussian waveforms, both with autocorrelation $\phi(\tau)$ given by

$$\phi(\tau) = \int_0^\infty Y(\omega) \cos (\omega - \omega_0)\tau \, d\omega \tag{2}$$

If the sine wave is now narrow-band modulated, so that an information-bearing waveform $x(t)$ is transmitted,

$$x(t) = x_s(t) \sin \omega_0 t + x_c(t) \cos \omega_0 t \tag{3}$$

The signal $z(t)$, observed following scatter, is then

$$z(t) = z_s(t) \sin \omega_0 t + z_c(t) \cos \omega_0 t \tag{4}$$

where

$$z_s(t) = x_s(t)\, y_s(t) - x_c(t)\, y_c(t) \tag{5}$$

and

$$z_c(t) = x_s(t)\, y_c(t) + x_c(t)\, y_s(t) \tag{6}$$

Thus scatter may be pictured as a complex multiplicative process.

To make the channel model as realistic as possible, we must not neglect the presence of additive shot- and thermal-noise. We shall consider this gaussian noise $n(t)$ to be localized at the receiver input, and shall assume that its spectrum $N(\omega)$ is symmetric and narrow-band about the carrier frequency ω_0, although broad relative to the spectrum of $z(t)$.

$$n(t) = n_s(t) \sin \omega_0 t + n_c(t) \cos \omega_0 t \tag{7}$$

$n_s(t)$ and $n_c(t)$ are independent gaussian waveforms, both with autocorrelation $\phi_n(\tau)$ given by

$$\phi_n(\tau) = \int_0^\infty N(\omega) \cos (\omega - \omega_0)\tau \, d\omega \tag{8}$$

Combining scatter and noise, the signal $w(t)$ finally observed at the receiver is

$$w(t) = w_s(t) \sin \omega_0 t + w_c(t) \cos \omega_0 t \tag{9}$$

where

$$w_s(t) = x_s(t)\, y_s(t) - x_c(t)\, y_c(t) + n_s(t) \tag{10}$$

and

$$w_c(t) = x_s(t)\, y_c(t) + x_c(t)\, y_s(t) + n_c(t) \tag{11}$$

Since $y_s(t)$, $y_c(t)$, $n_s(t)$ and $n_c(t)$ are independent gaussian waveforms, $w_s(t)$ and $w_c(t)$ share a joint gaussian distribution, $p\left[w_s(t), w_c(t)/x_s(t), x_c(t)\right]$, assuming that $x(t)$ is known <u>apriori</u>.

B. The Probability-Computing Receiver

As mentioned in the introduction, the transmitter is assumed to conform to the Shannon model; that is, it contains an information source which generates a sequence of symbols drawn randomly and independently from a finite alphabet of size M. These symbols are encoded one-to-one into voltage waveforms $x^{(k)}(t)$, $k = 1, 2, \ldots, M$, all of duration T and having the character of modulated carriers, as specified in Section A. The receiver cannot know the x(t) sequence *apriori* if any information is to be conveyed, and the channel perturbations are such that observation of the received signal is not sufficient to state with certainty which symbols were transmitted. Under these conditions, the best which a receiver can do, within the appropriate mathematical framework of statistics, is to calculate the conditional probabilities $P\left[x^{(k)}(t)/w(t)\right]$, $k = 1, 2, \ldots, M$, symbol-by-symbol.

Using Baye's Theorem, and assuming that the *apriori* $P\left[x^{(k)}(t)\right]$ are all equal, we have

$$P\left[x^{(k)}(t)/w(t)\right] = K(w)\, p\left[w_s(t), w_c(t)/x^{(k)}(t)\right] \tag{12}$$

where K(w) is constant for all k, for any given symbol interval. The probability computation is generally simplified if $w_s(t)$ and $w_c(t)$ are conditionally independent, so that

$$p\left[w_s(t), w_c(t)/x^{(k)}(t)\right] = p\left[w_s(t)/x^{(k)}(t)\right]\; p\left[w_c(t)/x^{(k)}(t)\right] \tag{13}$$

In general, $\overline{w_s(t_1)w_c(t_2)}\Big|_{x^{(k)}(t)} \neq 0$, so that this cannot be.

In order to achieve conditional independence for general transmissions, it is necessary to seek two new variables f(t) and g(t) through a linear transformation of $w_s(t)$ and $w_c(t)$. If we assume that the noise has uniform spectral density in its bandwidth B, and that the waveforms f(t) and g(t) are sampled only at intervals of 1/B, there exists a family of transformations which will give the desired conditional independence. The key transformation involves $x^{(k)}(t)$ itself:

$$\left.\begin{aligned} f^{(k)}(t) &= \frac{x_s^{(k)}(t)}{r^{(k)}(t)}\; w_s(t) + \frac{x_c^{(k)}(t)}{r^{(k)}(t)}\; w_c(t) \\ g^{(k)}(t) &= \frac{x_s^{(k)}(t)}{r^{(k)}(t)}\; w_c(t) - \frac{x_c^{(k)}(t)}{r^{(k)}(t)}\; w_s(t) \\ r^{(k)}(t) &= \sqrt{\left[x_s^{(k)}(t)\right]^2 + \left[x_c^{(k)}(t)\right]^2} \end{aligned}\right\} \tag{14}$$

Any further "rotations" will leave conditional independence unaffected:

$$\left.\begin{aligned} f^{*(k)}(t) &= a f^{(k)}(t) + b g^{(k)}(t) \\ g^{*(k)}(t) &= a g^{(k)}(t) - b f^{(k)}(t) \\ a^2 &+ b^2 = 1 \end{aligned}\right\} \tag{15}$$

We are now in a position to give detailed expressions which approximate the $p\left[f^{(k)}(t),\ g^{(k)}(t)/x^{(k)}(t)\right]$ from observation of $x^{(k)}(t)$ and $w(t)$ at their sampling points only. Letting the sampled independent pairs $f_i^{(k)}, g_j^{(k)}$ or $f_i^{*(k)}, g_j^{*(k)}$ be denoted by $u_i^{(k)}$ and $v_j^{(k)}$, we have, from the multiple-order gaussian distribution,[10]

$$p\left[u_1^{(k)},u_2^{(k)},\ldots,u_n^{(k)};\ v_1^{(k)},v_2^{(k)},\ldots,v_n^{(k)}/x^{(k)}(t)\right] \tag{16}$$

$$= (2\pi)^{-n}\left|M_n^{(k)}\right|^{-1}\ \exp\left\{-1/2\ \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{M_n^{(k)ij}}{\left|M_n^{(k)}\right|}\left[u_i^{(k)}u_j^{(k)}+v_i^{(k)}v_j^{(k)}\right]\right\}$$

$M_n^{(k)ij}$ is the cofactor of $m_{ij}^{(k)} = \overline{u_i^{(k)}u_j^{(k)}}\Big|_{x^{(k)}(t)} = \overline{v_i^{(k)}v_j^{(k)}}\Big|_{x^{(k)}(t)}$ in the matrix $M_n^{(k)}$,

and $\left|M_n^{(k)}\right|$ is its determinent.

$$M_n^{(k)} = \begin{bmatrix} m_{11}^{(k)} & m_{12}^{(k)} & \cdots & m_{1n}^{(k)} \\ m_{12}^{(k)} & m_{22}^{(k)} & \cdots & m_{2n}^{(k)} \\ \cdot & \cdot & \cdots & \cdot \\ m_{1n}^{(k)} & m_{2n}^{(k)} & \cdots & m_{nn}^{(k)} \end{bmatrix}, \quad n = BT \tag{17}$$

We find,

$$m_{ij}^{(k)} = r_i^{(k)}\, r_j^{(k)}\ \phi_{ij} + N\delta_{ij} \tag{18}$$

where $\phi_{ij} = \phi\left[(i-j)/B\right]$, N is the noise power, and δ_{ij} is the Kronecker delta-function: $\delta_{ii} = 1$; $\delta_{ij} = 0$, $i \neq j$. From (14), $r_i^{(k)}$ is the amplitude of the envelope of $x(t)$ at the i-th sampling point.

A simplification results if $x(t)$ is an amplitude-modulated carrier: $x_c(t) = C\, x_s(t)$. Then, from (14),

$$u_i^{(k)}u_j^{(k)} + v_i^{(k)}v_j^{(k)} = \left[w_{si}w_{sj} + w_{ci}w_{cj}\right]\ \frac{x_{si}^{(k)}\, x_{sj}^{(k)}}{\left|x_{si}^{(k)}\ x_{sj}^{(k)}\right|} \tag{19}$$

Thus, in this case the transformation (14) is not necessary to achieve conditional independence. The ± 1 factor on the right side of (19) may easily be included in the $M_n^{(k)ij}$.

After the $p\left[u,v/x^{(k)}(t)\right]$ have been computed by (16), the direct probabilities $P\left[x^{(k)}(t)/w(t)\right]$ follow from Baye's Theorem, as in (12). A schematic diagram of the overall operation of the probability-computing receiver appears in Figure 1. The one approximation made in the preceding analysis has been the neglect of possible information present in the waveforms at other than sampling points. This difference, however, becomes negligibly small as the points are taken closer together by letting the noise bandwidth B increase.

C. Receiver Simplification at Small Signal-to-Noise Ratios

In general, the inversion of the high-order matrix $M_n^{(k)}$, required to obtain the $M_n^{(k)ij}$, is a tedious process. In the limit as $n \to \infty$ for a fixed length of observation T, the sampling points become so dense that all the information is extracted. Inversion of $M_n^{(k)}$ then involves the solution of an integral equation. While such infinite-order matrices have successfully been inverted, such as those considered by Reich and Swerling,[4] the $m_{ij}^{(k)}$ considered here apparently do not yield a tractable solution. It has been possible, however, to obtain an approximation to the ideal receiver for small signal-to-noise ratios which permits considerable simplification of operation.

Let us assume, as before, that the additive gaussian noise has power N and is uniform over the bandwidth B, and that the sample points are taken 1/B apart. We shall consider the case of general transmission, and hence deal with $f^{(k)}(t)$ and $g^{(k)}(t)$, obtained from (14). From (10) and (11) it is easily shown that

$$p\left[\text{all } f_i^{(k)}, g_i^{(k)}/\text{all } x_{si}^{(k)}, x_{ci}^{(k)}, y_{si}, y_{ci}\right] = \tag{20}$$

$$(2\pi N)^{-n} \exp\left\{-\frac{1}{2N}\sum_{i=1}^{n}\left[\left(f_i^{(k)} - y_{si}r_i^{(k)}\right)^2 + \left(g_i^{(k)} - y_{ci}r_i^{(k)}\right)^2\right]\right\}$$

Expanding the squares, we obtain linear and square terms in y_{si} and y_{ci}. Analysis of the variances of these terms shows that the contributions of the square terms are negligible compared to the linear, providing the signal-to-noise ratio in the band occupied by the signal is small compared to unity. Thus,

$$p\left[\text{all } f_i^{(k)}, g_i^{(k)}/\text{all } x_{si}^{(k)}, x_{ci}^{(k)}, y_{si}, y_{ci}\right] \approx \tag{21}$$

$$\exp\left\{-\frac{1}{2N}\sum_{i=1}^{n}\left(\left[f_i^{(k)}\right]^2 + \left[g_i^{(k)}\right]^2\right)\right\} \exp\left\{\frac{1}{N}\sum_{i=1}^{n}\left[f_i^{(k)}r_i^{(k)}y_{si} + g_i^{(k)}r_i^{(k)}y_{ci}\right]\right\}$$

where $\approx$ indicates "approximately proportional to". In order to obtain $p\left[\text{all } f_i^{(k)}, g_i^{(k)}/\text{all } x_{si}^{(k)}, x_{ci}^{(k)}\right]$ from (21), we must average (21) over the gaussianly-distributed y_{si} and y_{ci}. Recognizing the similarity of this exponential average to the characteristic function, and using the result given by Rice[10] for the gaussian case, we obtain

$$p\left[\text{all } f_i^{(k)}, g_i^{(k)} / \text{all } x_{si}^{(k)}, x_{ci}^{(k)}\right] \tilde{\propto} \tag{22}$$

$$\exp\left\{ -\frac{1}{2N} \sum_{i=1}^{n} \left(\left[f_i^{(k)}\right]^2 + \left[g_i^{(k)}\right]^2 \right) \right\} \exp\left\{ \frac{1}{2N^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[f_i^{(k)} f_j^{(k)} + g_i^{(k)} g_j^{(k)} \right] r_i^{(k)} r_j^{(k)} \phi_{ij} \right\}$$

in the limit as $n \to \infty$,

$$p\left[f^{(k)}(t), g^{(k)}(t) / x^{(k)}(t)\right] \tilde{\propto} \exp\left\{ \frac{1}{2N^2 \Delta^2} \left[\int_0^T \int_0^T f^{(k)}(t_1) f^{(k)}(t_2) r^{(k)}(t_1) r^{(k)}(t_2) \phi(t_1 - t_2) dt_1\, dt_2 + \int_0^T \int_0^T g^{(k)}(t_1) g^{(k)}(t_2) r^{(k)}(t_1) r^{(k)}(t_2) \phi(t_1 - t_2) dt_1\, dt_2 \right] \right\} \tag{23}$$

where $\Delta = 1/B$ is the sampling interval. Eq. (23) applies equally well to the $f^{*(k)}(t)$ and $g^{*(k)}(t)$ of (15). When x(t) is an amplitude-modulated carrier, a relation similar to (19) yields

$$p\left[f^{(k)}(t), g^{(k)}(t) / x^{(k)}(t)\right] \tilde{\propto} \tag{24}$$

$$\exp\left\{ \frac{1}{2N^2 \Delta^2} \left[\int_0^T \int_0^T w_s(t_1) w_s(t_2) R^{(k)}(t_1) R^{(k)}(t_2) \phi(t_1 - t_2) dt_1\, dt_2 + \int_0^T \int_0^T w_c(t_1) w_c(t_2) R^{(k)}(t_1) R^{(k)}(t_2) \phi(t_1 - t_2) dt_1\, dt_2 \right] \right\}$$

where $R^{(k)}(t)$ has the magnitude of $r^{(k)}(t)$ and the sign of $x^{(k)}(t)$. Thus for the AM case, transformation (14) is not necessary.

Evaluation of the integrals in (23) and (24) could be performed by an analogue device using linear filters and multipliers. We invoke the fact that if F(s,t) is a symmetric function, that is, F(s, t) = F(t,s), in the range $0 \le t \le T$ and $0 \le s \le T$, then

$$\int_0^T \int_0^T F(s,t)\, ds\, dt = 2 \int_0^T \int_0^S F(s,t)\, dt\, ds \tag{25}$$

By inspection, the integrals of (23) and (24) satisfy the conditions, for $\phi(t-s) = \phi(s-t)$. Thus, for example, the first integral of (24) becomes

$$\int_0^T w_s(t_1) R^{(k)}(t_1) \int_0^{t_1} w_s(t_2) R^{(k)}(t_2) \phi(t_1 - t_2)\, dt_2\, dt_1$$

The analogue device would multiply $w_s(t)$ by $R^{(k)}(t)$, pass the product through a filter with impulse response $h(t) = \phi(t), t \geq 0$, form a second product using the first and filtered products, and finally pass the second product into an integrating filter. The output of the integrating filter at time T is then the value of the integral. A schematic diagram of the operation appears in Fig. 2.

Transformation (14) may also be performed by analogue. Let $x^{(k)}(t)$ be clipped and filtered to form a new waveform $X^{(k)}(t)$ having unit amplitude and preserving the phase modulation of $x^{(k)}(t)$. Then if the double-carrier-frequency terms are filtered from the product $w(t)X^{(k)}(t)$, $f^{(k)}(t)$ is obtained. Similarly, multiplication of $w(t)$ by $X^{(k)}(t)$ after the latter has been passed through a ninety-degree phase-shift network yields $g^{(k)}(t)$. It is not necessary to preserve absolute phase in the $x^{(k)}(t)$ stored at the receiver, for phase shift merely performs the transformation (15), to which the probability expressions are invariant.

Conclusions

A new type of channel disturbance, having a multiplicative nature, has been analyzed, and the appropriate ideal receiver has been synthesized. It has further been possible to simplify the receiver to elementary analogue operations, providing the signal-to-noise ratio is sufficiently small. The success of the analysis rests almost wholly on the elegant properties of the gaussian distribution function, and it is not expected that generalization to other multiplicative disturbances could be accomplished. Application of the results of this paper is an open question, but construction of a laboratory model of the transmitter, channel and receiver might prove interesting.

References

1. C. E. Shannon and W. Weaver, The Mathematical Theory of Communication, University of Illinois (1949)
2. P. M. Woodward and I. L. Davies, Information Theory and Inverse Probability in Telecommunications, Proc. IEE, III, p. 37 (March 1952)
3. J. H. Van Vleck and D. Middleton, Jour. Appl. Phys. 17, 940 (1946)
4. E. Reich and P. Swerling, The Detection of a Sine Wave in Gaussian Noise, Jour. Appl. Phys. 24,289 (1953)
5. H. G. Booker and W. E. Gordon, A Theory of Radio Scattering in the Troposphere, Proc.IRE 38,401 (1950)
6. H. G. Booker, J. A. Ratcliffe, D. H. Shinn, Diffraction from a Random Screen with Applications to Ionospheric Problems, Phil. Trans. 242,579 (1950).
7. S. O. Rice, Statistical Fluctuations of Radio Field Strength Far Beyond the Horizon, Proc. IRE 41,274 (1953)
8. Lord Rayleigh, Phil. Mag. (6) 37,321 (1919). Also Scientific Papers, 6,604 (1920)
9. S. O. Rice, Mathematical Analysis of Random Noise, BSTJ, 24, 46-156, Sec. 3.7 (1945)
10. S. O. Rice, Mathematical Analysis of Random Noise, BSTJ, 23, 282-332, Sec. 2.9 (1944).

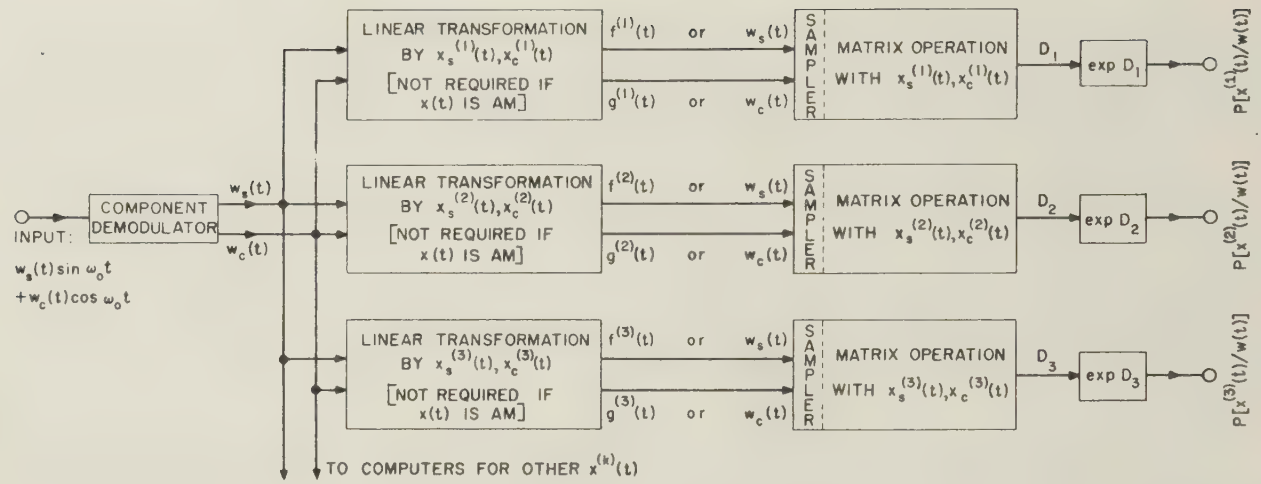


Fig. 1 - Probability-computing receiver for scatter channel.

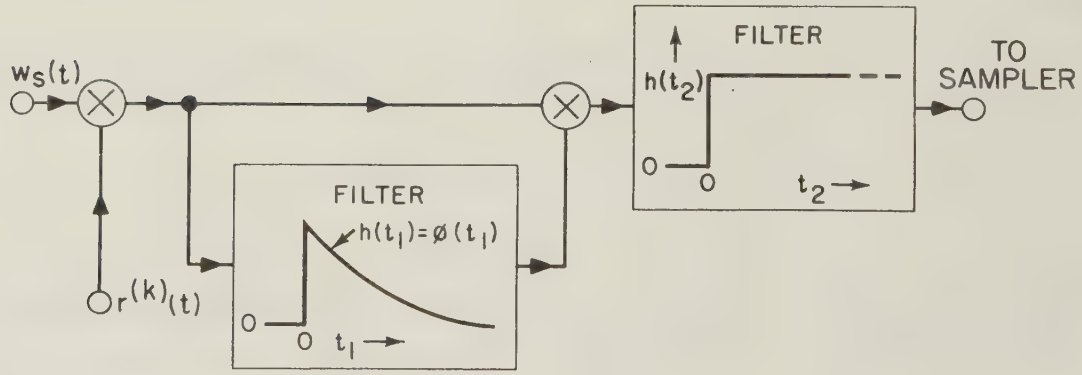


Fig. 2 - Element of approximate probability-computing receiver.

THE THEORY OF SIGNAL DECTECTABILITY *

W. W. Peterson, T. G. Birdsall, and W. C. Fox
University of Michigan
Ann Arbor, Michigan

ABSTRACT

The problem of signal detectability treated in this paper is the following: Suppose an observer is given a voltage varying with time during a prescribed observation interval and is asked to decide whether its source is noise or is signal plus noise. What method should the observer use to make this decision, and what receiver is a realization of that method? After giving a discussion of theoretical aspects of this problem, the paper presents specific derivations of the optimum receiver for a number of cases of practical interest.

The receiver whose output is the value of the likelihood ratio of the input voltage over the observation interval is the answer to the second question no matter which of the various optimum methods current in the literature is employed including the Neyman - Pearson observer, Siegert's ideal observer, and Woodward and Davies' "observer." An optimum observer required to give a yes or no answer simply chooses an operating level and concludes that the receiver input arose from signal plus noise only when this level is exceeded by the output of his likelihood ratio receiver.

Associated with each such operating level are conditional probabilities that the answer is a false alarm and the conditional probability of detection. Graphs of these quantities, called receiver operating characteristic, or ROC, curves are convenient for evaluating a receiver. If the detection problem is changed by varying, for example, the signal power, then a family of ROC curves is generated. Such things as betting curves can easily be obtained from such a family. The operating level to be used in a particular situation must be chosen by the observer. His choice will depend on such factors as the permissable false alarm rate, a priori probabilities, and relative importance of errors.

With these theoretical aspects serving as an introduction, attention is devoted to the derivation of explicit formulas for likelihood ratio, and for probability of detection and probability of false alarm, for a number of particular cases. Stationary, band-limited, white Gaussian noise is assumed. The seven special cases which are presented were chosen from the simplest problems in signal detection which closely represent practical situations.

Two of the cases form a basis for the best available approximation to the important problem of finding probability of detection when the starting time of the signal, signal frequency, or both, are unknown. Furthermore, in these two cases uncertainty in the signal can be varied, and a quantitative relationship between uncertainty and ability to detect signals is presented for these two rather general cases. The variety of examples presented should serve to suggest methods for attacking other simple signal detection problems and to give insight into problems too complicated to allow a direct solution.

1. INTRODUCTION

The problem of signal detectability treated in this paper is that of determining a set of optimum instructions to be issued to an "observer" who is given a voltage varying with time during a prescribed observation interval and who must judge whether its source is "noise" or "signal plus noise." The nature of the "noise" and of the "signal plus noise" must be known to some extent by the observer.

Any equipment which the observer uses to make this judgement is called the "receiver." Therefore the voltage with which the observer is presented is called the "receiver input." The optimum instructions may consist primarily in specifying the "receiver" to be used by the observer.

The first three sections of this article survey the applications of statistical methods to this problem of signal detectability. They are intended to serve as an introduction to the subject to those who possess a minimum of mathematical training. Several definitions of "optimum" instructions have been proposed by other authors. Emphasis is placed here on the fact that these various definitions lead to essentially the same receiver. In subsequent sections the actual specification of the optimum receiver is carried out and its performance is evaluated numerically for some cases of practical interest.[17]

---

* The work reported in this paper was done under U.S. Army Signal Corps Contract No. DA - 36 - 039 sc - 15358.

1.1 Population SN and N

Either noise alone or the signal plus noise may be capable of producing many different receiver inputs. The totality of all possible receiver inputs when noise alone is present is called "Population N"; similarly, the collection of all receiver inputs when signal plus noise is present is called "Population SN." The observer is presented with a receiver input from one of the two populations, but he does not know from which population it came; indeed, he may not even know the probability that it arose from a particular population. The observer must judge from which population the receiver input came.

1.2 Sampling Plans[1]

A sampling plan is a system of making a sequence of measurements on the receiver input during the observation interval in such a way that it is possible to reconstruct the receiver input for the observation interval from the measurements. Mathematically, a sampling plan is a way of representing functions of time as sequences of numbers. The simplest way to describe this idea is to list a few examples.

A: Fourier Series on an Interval Suppose that the observation interval begins at time t_o and is T seconds long, and that each function in the population SN and N can be expanded in a Fourier series on the observation interval. The Fourier coefficients for each particular receiver input can be obtained by making measurements on that input, which can in turn be reconstructed from these measurements by the formula

$$x(t) = a_o + \sum_{n=1}^{\infty} a_n \cos \frac{2\pi nt}{T} + b_n \sin \frac{2\pi nt}{T}, \quad t_o < t < t_o + T. \tag{1}$$

Thus the process representing each function x(t) by the sequence of its Fourier coefficients (a_o, a_1, b_1, ... , a_n, b_n, ...) is a sampling plan in the sense described above.

The pair of terms in the Fourier series which involve the cosine and sine of $2\pi nt/T$ is of frequency n/T cycles per second. Suppose that for a particular population of receiver inputs the terms of frequency greater than n_o/T are zero; i.e., the population is bandlimited in the Fourier series sense or simply "series-bandlimited." For such a population the process of representing each receiver input x(t) by the finite sequence (a_o, a_1, b_1, ... , a_{n_o}, b_{n_o}) is a finite sample plan.*

B: Shannon's Sampling Plan Suppose that the observation interval includes all time and that the populations are "transform-bandlimited" to a band from 0 to W cycles per second, i.e., the Fourier transform of every receiver input is zero for frequencies greater than W. A sampling plan for this population is to represent each function x(t) by its amplitude measured at times spaced 1/2W seconds apart, (... $x(t_o-n/2W)$, ... , $x(t_o-1/2W)$, $x(t_o)$, $x(t_o+1/2W)$, ... $x(t_o+n/2W)$, ...). In this case the formula[2] for the reconstruction of the receiver input is

$$x(t) = \sum_{n=-\infty}^{\infty} x(t_o + \frac{n}{2W}) \frac{\sin \pi (2W(t-t_o) - n)}{\pi (2W(t-t_o) - n)} \cdot \tag{2}$$

The instants of time $t_o+n/2W$ are called sampling-times. Each choice of t_o between 0 and 1/2W yields a different sampling plan. If the observation interval again includes all time, but the populations are transform-bandlimited to a frequency band from $f_o-W/2$ to $f_o+W/2$ which does not contain zero frequency, then each receiver input x(t) can be considered as an amplitude and frequency modulated waveform, $x(t)=r(t) \cos (2\pi f_o t + \theta(t))$; r(t) is the amplitude of the envelope and $\theta(t)$ is the instantaneous phase of the carrier. A sampling plan employing sampling-times is obtained in this case by representing each receiver input by the sequence (... $r(t_o)$, $\theta(t_o)$, ... , $r(t_o+n/W)$, $\theta(t_o+n/W)$, ...) of envelope amplitudes and carrier phases measured at sampling-times spaced by 1/W seconds apart.[1] The reconstruction of the receiver input from this sequence is given by

$$x(t) = \sum_{n=-\infty}^{\infty} r(t_o+\frac{n}{W}) \cdot \cos\left(2\pi t_o t + \theta(t_o+\frac{n}{W})\right) \frac{\sin \pi (W(t-t_o)-n)}{\pi (W(t-t_o) - n)} \cdot \tag{3}$$

C: Sampling Plan Using Sampling-Times for a Finite Observation Interval Only functions known for all times have Fourier transforms, and therefore the hypothesis that the populations are transform-bandlimited applies only when the observation interval includes all time. If the observation interval is of finite length and if the populations are series-bandlimited, then there are sampling

---

* A sampling plan is finite if there is a finite maximum length for the sequences for all receiver inputs in the population.

plans utilizing sampling-times which are similar to those described in paragraph B for transform-band-limited populations and an infinite observation interval. Suppose that time is measured from the beginning of the observation interval, which is T seconds long, and suppose that the populations are series-bandlimited from 0 to W cycles per second. A finite sampling plan for this situation can be obtained by representing each receiver input by the sequence of its amplitudes measured 1/2W seconds apart,[1]

$$\left(x(t_o),\ x(t_o + \frac{1}{2W}),\ \ldots\ ,\ x(t_o + T - \frac{1}{2W})\right) \tag{4}$$

and the reconstruction of the receiver input from this sequence is

$$x(t) = \sum_{n=0}^{2WT-1} x(t_o + \frac{n}{2W}) \quad \frac{\sin \pi (2W(t-t_o) - n)}{2WT \sin\left(\frac{2W(t-t_o) - n}{2WT}\pi\right)}, \quad 0 < t < T\ . \tag{5}$$

Again each choice of the (initial) sampling-time t_o between 0 and 1/2W yields a different sampling plan. In a similar fashion, if the observation interval is unchanged but the populations are series-band-limited on this interval to a frequency band from f_o-W/2 to f_o+W/2 which does not include zero frequency, then each receiver input can be represented by a finite sequence $(r(t_o), \theta(t_o), r(t_o+1/W), \theta(t_o+1/W), \ldots, r(t_o+T-1/W), \theta(t_o+T-1/W))$ of envelope amplitudes and carrier phases measured at sample points 1/W seconds apart; t_o is again used to denote the initial sampling time which may be chosen anywhere from 0 to 1/W. The reconstruction of the receiver input from this sequence of measurements is given by

$$x(t) = \sum_{n=0}^{WT-1} r(t_o+\frac{n}{W}) \cos\left(2\pi f_o t + \theta(t_o+\frac{n}{W})\right) \frac{\sin \pi\ (W(t-t_o) - n)}{WT \sin \pi\ \frac{W(t-t_o) - n}{WT}}\ , \quad 0 < t < T\ . \tag{6}$$

From these examples it can be seen that there are a number of important differences between various sampling plans such as i) the length of the observation interval, ii) whether sampling-times are employed, and iii) whether the measurements are all to be of the same kind, e.g., instantaneous amplitude measurements, or of different kinds, e.g., envelope amplitude and carrier phase. However, they all have in common the property that the receiver input can be reconstructed from the measurements made on it.

The role which the sampling plan plays in the theory presented in this paper is primarily one of mathematical convenience. The populations N and SN will be represented as sequences through the use of sampling plans in order to apply statistical methods. Once an answer is obtained concerning an "optimum" receiver, it is often possible to translate this answer back to the more familiar language of receiver inputs. If a finite sampling plan is not available for a particular application of the theory, then recent work by Grenander[3] shows that the desired parameters of the "optimum" receiver can be approximated by using finite sampling plans. Both for this reason and in order to simplify the exposition, the theory presented here is restricted to cases where finite sampling plans are available.

2. OPTIMUM TESTS ON FIXED OBSERVATION INTERVALS

2.1 Probability Density Functions

This part of the paper is concerned with a method of statistical analysis which requires for raw data a finite sequence of numbers $(x_1, x_2, \ldots, x_n)$, which is the result of the measurements made at the receiver input according to some particular finite sampling plan. The sequence is often called a "sample" of the population from which it arose, and is denoted by a single letter; thus, if the receiver input is x(t), and the sampling plan yields a sequence $(x_1, x_2, \ldots, x_n)$, then this sequence is called the sample X. The theory to be developed here is intended to specify an optimum receiver and is couched in the language of samples, $X = (x_1, x_2, \ldots, x_n)$. If n is very large, a receiver which had to make the measurements called for by a sampling plan would certainly be impractical. However, this practical difficulty is avoided when the specification of the receiver is translated back from the language of samples to the language of the receiver inputs; this can be done because it is possible to reconstruct the inputs from the samples.

For the purposes of the subsequent development any finite sampling plan may be considered provided

enough properties are known of the associated sample X so that certain probabilities may be calculated. Specifically, the probability density functions $f_N(X)$ and $f_{SN}(X)$ of the sample variable X for the cases when X is drawn from populations N and SN respectively must be known.* The two basic properties of density functions are

$$f_N(X) \geq 0 \qquad \int f_N(X)\, dX = 1, \tag{7}$$

and

$$f_{SN}(X) \geq 0 \qquad \int f_{SN}(X)\, dX = 1$$

where the integration symbol represents the multiple integral taken over the entire range of the sample variable $X = (x_1, x_2, \ldots, x_n)$.

2.2 The Concept of a Criterion

Consider now an observer who has as available data the sample $X = (x_1, \ldots, x_n)$. The observer's job is to judge for each sample whether or not it was taken from population SN. Although it is not possible to determine the (probably subconscious) criterion used by the observer, it is quite possible to find an external manifestation of it. Ideally all that is necessary is to submit each possible sample to the observer and to record his judgement. This will yield a tabulation of those samples which the observer decided were drawn from population SN. If any other observer is given this tabulation and instructed to base his decisions on it, he will behave exactly as did the first observer. Thus, the tabulation of these responses can be used to replace the mental criterion employed by the observer. Such a tabulation will also be called a criterion and will be denoted by the letter A, which refers to the phraseology common in statistics of "Accepting the hypothesis that a signal is present." The tabulation of the remaining samples, those which the observer concluded were drawn from population N, will be denoted by B.

2.3 Probabilities Associated with Criteria

There are of course as many different criteria as there are observers. Among all possible criteria it is necessary to select those that are best for various purposes. To do so, certain numerical quantities must be associated with each criterion. It will be necessary to know the probability that a sample from one of the populations will be listed in a particular criterion A. According to the standard definitions, these probabilities are given by

$$P_{SN}(A) = \int f_{SN}(X)\, dX$$

and (8)

$$P_N(A) = \int f_N(X)\, dX$$

where the multiple integral is taken over all samples listed in the criterion A.

For example, a particular sample plan might have a density function of the form $f_N(x_1, x_2, \ldots, x_n) = K \exp -(x_1^2 + x_2^2 + \ldots + x_n^2)$. A possible criterion would consist of those samples $X = (x_1, x_2, \ldots, x_n)$ which lie outside a sphere of radius one centered at the origin. Then the integral would be taken over the exterior of this sphere.

These probabilities have a special significance. $P_N(A)$ is the conditional probability that a sample from population N will be listed in criterion A, that is, will be judged as a sample from population SN. Thus $P_N(A) = F$ is the conditional false alarm probability. Also, $P_{SN}(A)$ is the conditional probability of a certain kind of correct response called a hit (that of judging correctly that a sample is from population SN). The conditional probability of judging falsely that a sample is from population SN is therefore given by $1 - P_{SN}(A) = M$, the conditional probability of a miss. The only errors which can occur are false alarms and misses; their conditional probabilities, F and M, are called briefly the error probabilities.

A reader familiar with the formal content of probability theory should note that these quantities

---

* In this discussion it should be kept in mind that "the event of the sample being drawn from population SN" corresponds to signal and noise being present at the receiver input. Also "the event of population SN being sampled" means the same thing.

are true conditional probabilities; the first is conditional on the sample being drawn from population SN; the second is conditional on its being drawn from population N. This is to distinguish them from a priori probabilities (the probabilities that a certain population will be sampled, for example) which are not as yet assumed known.

2.4 Likelihood Ratio and the Ratio Criteria

It is convenient to introduce a new function called the likelihood ratio, $\ell(X)$, defined as the ratio $f_{SN}(X)/f_N(X)$ for sample points $X = (x_1, \ldots, x_n)$; $\ell(X)$ represents the likelihood that the sample X was drawn from SN relative to the likelihood that it was drawn from N. Hence, if $\ell(X)$ is sufficiently large, it would be reasonable to conclude that X was in fact drawn from population SN, i.e., that X should be listed in the desired "best" criterion. Thus, for each number $\beta \geq 0$, a certain criterion A (β) will be selected; A(β) is chosen by listing each sample X for which $\ell(X) \geq \beta$. The problem then reduces to that of making a wise choice of β; that is, to determine how large "sufficiently large" is. Criteria of the form A(β) will be called ratio criteria.

A number of writers have presented varying definitions of a criterion being "optimum." It turns out that each of these optimum criteria can be expressed as a ratio criterion, so that a receiver designed to yield likelihood ratio as output could be used with any of them.

2.5 Weighted Combination Criteria

Suppose it is possible to assign a certain number w as a weighting factor representing the importance of a false alarm relative to a hit. Since $P_{SN}(A)$ is the probability of a hit, and $P_N(A)$ the probability of a false alarm, it would then be reasonable to find a criterion A which maximizes the quantity

$$P_{SN}(A) - wP_N(A). \tag{9}$$

But this quantity can be written as

$$\int_A \left[f_{SN}(X) - wf_N(X)\right] \, dX \tag{10}$$

where the integration is taken over the sample points X listed in A. To maximize this integral, one would list in A every sample for which the integrand was not negative. Solving that inequality for w, one sees that A should contain those sample points X for which

$$\ell(X) = \frac{f_{SN}(X)}{f_N(X)} \geq w. \tag{11}$$

Thus the desired criterion A is simply A(w), and so it is a ratio criterion.

2.6 Neyman-Pearson Criteria

If it is critically important to keep the probability of a false alarm $P_N(A)$ below a certain level k, then it would be reasonable to choose from among such criteria that one which maximizes the probability of a hit. Thus Neyman and Pearson proposed as a type of optimum criterion any criterion A_k for which

(1) $P_N(A_k) \leq k$, and
(2) $P_{SN}(A_k)$ is a maximum for all the criteria A with the property $P_N(A) \leq k$.

The A_k type criterion can also be expressed as a ratio criterion. This can be made plausible as follows. To begin with, it is necessary to consider only those criteria A for which $P_N(A) = k$, because A will be taken as large as possible in order to meet condition (2). Now consider the curve given parametrically by the equations

$$X = X(\beta) = P_N\,(A(\beta))$$

and

$$Y = Y(\beta) = P_{SN}\,(A(\beta)). \tag{12}$$

This curve will be called the Receiver Operating Characteristic (briefly, ROC) curve, for a receiver whose output is likelihood ratio and with which ratio criteria are being used.

The ROC curve passes through the points (0, 0) and (1, 1), the first at $\beta = \infty$, the second at $\beta = 0$. At $\beta = 0$, $\ell(X) \geq \beta = 0$ for all X, so A(0) consists of all possible samples, Thus the observer will report that every sample is drawn from SN, so he will be certain to make a false alarm and to make a hit. (This assumes that the samples will not be drawn exclusively from one of the populations.)

This can be verified, using the basic property of the density functions expressed by the following equations:

$$P_{SN}(A(0)) = \int f_{SN}(X)\, dX = 1$$

and (13)

$$P_N(A(0)) = \int f_N(X)\, dX = 1$$

where the integration is taken over all possible samples X. These equations mean that $X(0) = Y(0) = 1$. Moreover, $X(\infty) = Y(\infty) = 0$, because for $\beta = \infty$ there are no samples X with $\ell(X) \geq \infty$; i.e., $A(\infty)$ contains no samples at all and the operator will never report a signal is present. Therefore the operator cannot possibly make a false alarm nor can he make a hit. Thus $P_{SN}(A(\infty)) = 0$ and $P_N(A(\infty)) = 0$.

These considerations, together with those of the next section, show that the ROC curve can be sketched somewhat as in Fig. 1.

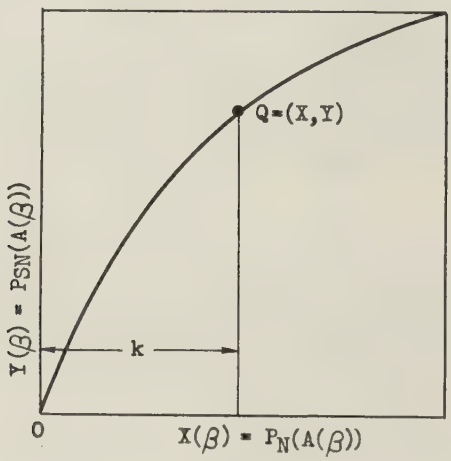


FIG. 1. TYPICAL ROC CURVE

To determine the desired A_k, recall that all probabilities lie between zero and one, so that $P_N(A_k) = k$ is between zero and one. Then there is a point Q of the ROC curve which lies vertically above the point (k,0). The coordinates (X,Y) of Q are $X = P_N(A(\beta)) = k$ and $Y = P_{SN}(A(\beta))$, for some β, which will be written β_k. Now $A(\beta_k)$ satisfies condition (1) because $P_N(A(\beta_k)) = k$, and therefore $A(\beta_k)$ will be the desired A_k if $P_{SN}(A) \leq P_{SN}(A(\beta_k))$ for any criterion with the property that $P_N(A) = k$. From paragraph 2.5, it is clear that the ratio criterion $A(\beta_k)$ is an optimum weighted-combination criterion with the weighting factor $w = \beta_k$. Therefore, if $w = \beta_k$, the weighted-combination using the criterion $A(\beta_k)$ is greater than or equal to the same weighted-combination using any other criterion A, i.e.,

$$P_{SN}(A(\beta_k)) - \beta_k P_N(A(\beta_k)) \geq P_{SN}(A) - \beta_k P_N(A) \tag{14}$$

In this case both $P_N(A(\beta_k))$ and $P_N(A)$ are equal to k. If this value is substituted into the inequality above, one obtains

$$P_{SN}(A(\beta_k)) \geq P_{SN}(A). \tag{15}$$

Therefore, the desired Neyman-Pearson criterion A_k should be chosen to be this particular ratio criterion, $A(\beta_k)$.

2.7 ROC Curve

It is desirable to digress for a moment to study the ROC curve more closely. Its value lies

in the fact that if the type of criterion chosen for a particular application is a ratio criterion, $A(\beta)$, then a complete description of the detection system's performance can be read off the ROC curve. By the very definition of the ROC curve, the X coordinate is the conditional probability, F, of false alarm, and the Y coordinate is the conditional probability of a hit. Similarly (1-X) is the conditional probability of being correct when noise alone is present, and (1-Y) = M is the conditional probability of a miss. It will be shown in a moment that the operating level β for the ratio criterion $A(\ )$ can also be determined from the ROC curve as the slope at the point

$$(P_N(A(\beta)),\ P_{SN}(A(\beta)))\ .$$

Since most proposed kinds of optimum criteria can be reduced to ratio criteria, the ROC curve assumes considerable importance.

In order to determine some of its geometric properties, it will be assumed that the parametric functions

$$X = X(\beta) = P_N(A(\beta))$$

and (16)

$$Y = Y(\beta) = P_{SN}(A(\beta))$$

are differentiable functions of β. The slope of the tangent to the ROC curve is given by the quotient $(dY/d\beta)/dX/d\beta$. To calculate the slope at the point $(X(\beta_o),\ Y(\beta_o))$, notice that among all criteria A, the quantity $P_{SN}(A) - \beta_o P_N(A)$ is maximized by $A = A(\beta_o)$. Therefore, in particular, the function

$$Y(\beta) - \beta_o X(\beta) = P_{SN}(A(\beta)) - \beta_o P_N(A(\beta)) \tag{17}$$

has a maximum at $\beta = \beta_o$, so that its derivative must vanish there. Thus differentiating,

$$\frac{dY}{d\beta} - \beta_o \frac{dX}{d\beta} = 0 \quad \text{at } \beta = \beta_o\ . \tag{18}$$

Solving for β_o, one obtains

$$\beta_o = \frac{\left(\frac{dY}{d\beta}\right)_{\beta = \beta_o}}{\left(\frac{dX}{d\beta}\right)_{\beta = \beta_o}} = \text{the slope of the tangent to the ROC curve at the point } (X(\beta_o),\ Y(\beta_o))\ . \tag{19}$$

This shows that the slope of the ROC curve is given by its parameter β, and so is always positive. Hence the curve rises steadily. In addition, this means that $Y(\beta)$ can be written as a single valued function of $X(\beta)$, $Y = Y(X)$, which is monotone increasing, and where $Y(0) = 0$ and $Y(1) = 1$. These remarks make fully warranted the sketch of the ROC curve given in Fig. 1. The next two sections are concerned with determining the best value to use for the weighting factor w when a priori probabilities are known.

2.8 Siegert's "Ideal Observer's" Criteria

Here it is necessary to know beforehand the a priori probabilities that population SN and that population N will be sampled. This is an additional assumption. These probabilities are denoted respectively by P(SN) and P(N). Moreover, P(SN) + P(N) = 1 because at least one of the populations must be sampled. The criterion associated with Siegert's Ideal Observer is usually defined as a criterion for which a priori probability of error is minimized (or, equivalently, the a priori probability of a correct response is maximized).[5] Frequently the only case considered is that where P(SN) and P(N) are equal, but this restriction is not necessary.

Since the conditional probability F of a false alarm is known as well as the a priori probability of the event (that population N was sampled) upon which F is conditional, then the probability of a false alarm is given by the product

$$P(N)F\ . \tag{20}$$

In the same way the probability of a miss is given by

$$P(SN)M\ . \tag{21}$$

Because an error E can occur in exactly these two ways, the probability of error is the sum of these quantities

$$P(E) = P(N)F + P(SN)M \quad . \tag{22}$$

It has already been pointed out that $F = P_N(A)$ and $M = 1-P_{SN}(A)$. If these are substituted into the expression for P(E) a simple algebraic manipulation gives

$$P(E) = P(SN) - P(SN)\left[P_{SN}(A) - \frac{P(N)}{P(SN)} \cdot P_N(A)\right] \quad . \tag{23}$$

It is desired to minimize P(E). But from the last equation this is equivalent to maximizing the quantity

$$P_{SN}(A) - \frac{P(N)}{P(SN)} \cdot P_N(A) \tag{24}$$

and, of course, this will yield a weighted combination criterion with $w = P(N)/P(SN)$, which is known to be simply a ratio criterion A(w).

2.9 Maximum Expected-Value Criteria

Another way to assign a weighting factor w depends on knowing the "expected value" of each criterion. This can be determined if the a priori probabilities P(SN) and P(N) are known, and if numerical values can be assigned to the four alternatives. Let V_D be the value of detection and V_Q the value of being "quiet", that is, of correctly deciding that noise alone is present. The other two alternatives are also assigned values, V_M, the value of a miss, and V_F, the value of a false alarm. The expected value associated with a criterion can now be determined. In this case it is natural to define an optimum criterion as one which maximizes the expected value. It can be shown that such a criterion maximizes

$$P_{SN}(A) - \left[\frac{P(N)}{P(SN)} \cdot \frac{V_Q - V_F}{V_D - V_M}\right] P_N(A) \tag{25}$$

By definition (see paragraph 2.5), this criterion is a weighted combination criterion with weighting factor

$$w = \frac{P(N)}{P(SN)} \cdot \frac{V_Q - V_F}{V_D - V_M} \tag{26}$$

and hence a likelihood ratio criterion. Seigert's "Ideal Observer" criterion is the special case for which $V_Q - V_F = V_D - V_M$.

2.10 A Posteriori Probability and Signal Detectability

Heretofore the observer has been limited to two possible answers, "signal plus noise is present" or "noise alone is present". Instead he may be asked what, to the best of his knowledge, is the probability that a signal is present. This approach has the advantage of getting more information from the receiving equipment. In fact, Woodward and Davies point out that if the observer makes the best possible estimate of this probability for each possible transmitted message, he is supplying all the information which his equipment can give him.[6] A good discussion of this approach is found in the original papers by Woodward and Davies.[6,7] Their formula for the a posteriori probability, $P_X(SN)$, becomes, in the notation of this paper,

$$P_X(SN) = \frac{f_{SN}(X)\ P(SN)}{f_{SN}(X)\ P(SN) + (1 - P(SN))\ f_N(X)} \quad , \quad \text{or} \tag{27}$$

$$P_X(SN) = \frac{\ell\ (X)\ P(SN)}{\ell\ (X)\ P(SN) + 1 - P(SN)} \tag{28}$$

If a receiver which has likelihood ratio as its output can be built, and if the a priori probability P(SN) is known, a posteriori probability can be calculated easily. The calculation could be built into the receiver calibration, since (28) is a monotonic function of $\ell(X)$; this would make the receiver an optimum receiver for obtaining a posteriori probability.

3. SEQUENTIAL TESTS WITH MINIMUM AVERAGE DURATION

3.1 Sequential Testing

The idea of sequential testing is this: make one measurement x_1 on the receiver input; if the evidence x_1 is sufficiently persuading, decide as to whether the receiver input was drawn from population SN or from population N. If the evidence is not so strong, make a second measurement x_2 and consider the evidence (x_1, x_2). Continue to make measurements until the resulting sequence of measurements is sufficiently pursuading in favor or one population or the other. Obviously this involves the theoretical possibility of making arbitrarily many measurements before a final decision is made. This does not mean that infinitely many measurements must be made in an actual application, nor does it necessarily mean that the operation might entail an arbitrarily long interval of time. If in a particular application measurements are taken at evenly spaced times then the "time base" of such a measurement plan is infinite. However, another plan might call for measurements to be made at the instants $t = 0$, $t = 1/2$, ... , $t = (n-1)/n$, ... and as these times all lie in the time interval from zero to one, such a measurement plan would have a time base of only one unit of time.

If the measurement plan has been carried out to the stage where n measurements $x_1, x_2, \dots, x_n$ have been made, the variable $X_n = (x_1, x_2, \dots, x_n)$ is called the n^{th} stage sample variable. A specific plan for measurements will be considered only if for each possible stage n, the two density functions $f_{SN}(X_n)$ and $f_N(X_n)$ of the n^{th} stage sample variable X_n are known; the first of these density functions is applicable when population SN is being sampled and the second is applicable when population N is being sampled. These density functions may very well differ at different stages, so that they should be written $f^n_N(X_n)$ and $f^n_{SN}(X_n)$; however, the n appearing in the argument X_n should always make the situation clear, and the superscript on the density functions themselves will be omitted.

3.2 Sequential Tests

A sequential test will consist of two things:

1) An (infinite) measurement plan with density functions $f_N(X_n)$ and $f_{SN}(X_n)$
2) An assignment of three criteria to each stage of the measurement plan.

These three criteria represent the three possible conclusions:

A) Signal plus noise is present, i.e. the sample comes from population SN
B) Noise alone is present, i.e. the sample comes from population N
C) Another measurement should be made.

At the first stage of the measurement plan, any (real) number at all could theoretically result from the first measurement. This means that the first stage sample variable $X_1 = (x_1)$ ranges through the entire number system, which will be written S_1 to stand for the first stage sample space. Suppose the three first-stage criteria A_1, B_1, and C_1, have been chosen. If the sample X_1 is listed in A_1, the conclusion that a signal is present is drawn and the test terminated. If it is listed in B_1 the conclusion is that noise alone is present, and again the test is terminated. If X_1 should be listed in C_1, another measurement will be made, and the test moves on to the second stage instead of terminating.

When the first stage criteria have been chosen, a limitation is placed on S_2, the space through which the second stage sample variable $X_2 = (x_1, x_2)$ ranges. The only way the test can proceed to the second stage is for $X_1 = (x_1)$ to be listed in C_1. Therefore, S_2 does not contain all possible second stage samples $X_2 = (x_1, x_2)$ but only those for which (x_1) is listed in C_1. Three second stage criteria, A_2, B_2, and C_2, must now be chosen from those samples X_2 listed in S_2. They must be chosen in such a way that there are no duplications in the listings and no sample in S_2 is omitted. These criteria carry exactly the same significance as those chosen in the first stage. That is, the three conclusions that a signal is or is not present, or that the test should be continued, are drawn when the sample X_2 is listed in A_2, B_2, or C_2 respectively.

The selection of criteria proceeds in the same way. If the n^{th} stage criteria A_n, B_n, and C_n, have been chosen, then the next stage's sample space S_{n+1} consists of those samples $X_{n+1} = (x_1, x_2, \dots, x_n, x_{n+1})$ for which $X_n = (x_1, x_2, \dots, x_n)$ was listed in C_n. Then from S_{n+1} are drawn the three (n+1) stage criteria A_{n+1}, B_{n+1}, and C_{n+1}.

When an entire sequence

$$\begin{aligned}&(A_1, B_1, C_1) \ ,\\ &(A_2, B_2, C_2) \ ,\\ &\quad\vdots\\ &(A_n, B_n, C_n) \ ,\\ &\quad\vdots\end{aligned}$$

of criteria is selected, a "sequential test" has been determined. This does not mean of course that the test will necessarily be particularly useful. However, among all the possible ways of selecting a sequence of criteria and hence a sequential test, there may be particular ones which are very useful.

3.3 Probabilities Associated with Sequential Tests

If Q_n is any n^{th} stage criterion, then the quantities*

and

$$\begin{aligned}P_N(Q_n) &= \int_{Q_n} f_N(X_n)\, dX_n\\ P_{SN}(Q_n) &= \int_{Q_n} f_{SN}(X_n)\, dX_n\end{aligned} \tag{29}$$

represent the (N or SN) conditional probabilities that an n^{th} stage sample X_n will be listed in the criterion Q_n. Conditional probabilities of particular interest are:

1) The n^{th} stage conditional error probabilities:
If population N is sampled, then the probability that the sample variable X_n will be listed in A_n is $P_N(A_n)$. This is the N-conditional probability of a false alarm.
If population SN is sampled, then the probability that the sample variable X_n will be listed in B_n is $P_{SN}(B_n)$. This is the SN-conditional probability of a miss.

2) The conditional error probabilities of the entire test:

$$F = \sum_{n=1}^{\infty} P_N(A_n), \text{ the N-conditional probability of a false alarm, and} \tag{30}$$

$$M = \sum_{n=1}^{\infty} P_{SN}(B_n), \text{ the SN-conditional probability of a miss,} \tag{31}$$

are merely the sums of the same error probabilities over all stages.

3) The conditional probabilities of terminating at stage n are

$$T_N^n = P_N(A_n) + P_N(B_n) \ , \text{ and} \tag{32}$$

$$T_{SN}^n = P_{SN}(A_n) + P_{SN}(B_n) \ . \tag{33}$$

These equations can be justified by a simple argument. The only way the test can terminate at stage n is for the sample variable X_n to be listed in either A_n or B_n. The probability of this event is the sum of the probabilities of the component events which are mutually exclusive since X_n can be listed in at most one of A_n and B_n.

---

* The notation $\int_{Q_n}$ indicates that the integration is to be carried out over all sample points listed in Q_n.

4) The conditional probabilities that the entire test will terminate are

$$T_N = \sum_{n=1}^{\infty} T_N^n \quad , \text{ and} \tag{34}$$

$$T_{SN} = \sum_{n=1}^{\infty} T_{SN}^n \quad . \tag{35}$$

3.4. Average Sample Numbers

There are two other quantities which must be introduced. One feature of the sequential test is that it affords an opportunity of arriving at a decision early in the sampling process when the data happens to be unusually convincing. Thus one might expect that, on the average, the stage of termination of a well-constructed sequential test would be lower than could be achieved by an otherwise equal, good standard test. It is therefore important to obtain expressions for the average or expected value of the stage of termination. As with other probabilities, there will be two of these quantities: one conditional on population N being sampled; the other conditional on population SN being sampled. They are given by

$$E_N = \sum_{n=1}^{\infty} n\, T_N^n \tag{36}$$

and

$$E_{SN} = \sum_{n=1}^{\infty} n\, T_{SN}^n \tag{37}$$

The letter E is used to refer to the term "expected value." The quantities E_N and E_{SN} are called the average sample numbers. The form these formulas take can be justified (somewhat freely) on the grounds that each value, n, which the variable "stage of termination" may take on must be weighted by the (conditional) probability that the variable will in fact take on that value.

It should be heavily emphasized that the average sample numbers are strictly average figures. In actual runs of a sequential test, the stages of termination will sometimes be less than the average sample numbers but will also be upon occasion much larger. Any sequential test whose average sample numbers are not finite would be useless for applications. Therefore the only ones to be considered are those with finite average sample numbers. Under this assumption,* it can be shown that $T_N = T_{SN} = 1$ so that the test is certain to terminate (in the sense of probability). On the other hand, if it is known that $T_N = T_{SN} = 1$ it does not always follow that the average sample numbers are finite. Such a situation would mean only that if a sequence of runs of the test were made, each run would probably terminate, but the average stage of termination would become arbitrarily large as more runs were made.

3.5 Sequential Ratio Tests

In studying non-sequential tests using finite samples it was found that the best criterion could always be expressed in terms of likelihood ratio. Therefore, it may be useful to introduce likelihood ratios at each stage of an infinite sample plan. The n^{th} stage likelihood ratio function $\ell(X_n)$ is defined as the ratio $f_{SN}(X_n)/f_N(X_n)$. Optimum criteria in the finite sample tests turned out to be criteria listing all samples X for which $\ell(X)$ is greater than or equal to a certain number. It should be possible to choose sequential criteria (A_n, B_n, C_n) in the same way. For each stage two numbers a_n and b_n with $b_n \leq a_n$ could be chosen. Then the criteria (A_n, B_n, C_n) determined by the numbers a_n and b_n would be

A_n lists all samples X_n of the sample space S_n for which $\ell(X_n) \geq a_n$
B_n lists all samples X_n of the sample space S_n for which $\ell(X_n) \leq b_n$
C_n lists all samples X_n of the sample space S_n for which $b_n < \ell(X_n) < a_n$.

If criteria selected in this way meet the requirements that the average sample numbers be finite, then the resulting sequential test is called a "sequential ratio test."

3.6 Optimum Sequential Tests

---

* Remember that the sampling process is not assumed to yield independence among the X_i.

It is customary[8] to define an optimum sequential test as that one for which the average sample numbers E_N and E_{SN} are minimum among all sequential tests with fixed error probabilities F and M.

In addition to the formulas given in Section 3.4, alternative formulas[9] for the average sample numbers are

$$E_N = 1 + \sum_{i=1}^{\infty} P_N(C_i) \tag{38}$$

and

$$E_{SN} = 1 + \sum_{i=1}^{\infty} P_{SN}(C_i) \, . \tag{39}$$

Thus, if a set of sequential criteria (A_n^*, B_n^*, C_n^*) is presented as a possible optimum test, then its optimum character is decided by ascertaining whether the inequalities

$$\sum P_N(C_i^*) \leq \sum P_N(C_i) \tag{40}$$

and

$$\sum P_{SN}(C_i^*) \leq \sum P_{SN}(C_i) \tag{41}$$

hold for every other set of sequential criteria $\{(A_n, B_n, C_n)\}$ with the same error probabilities, i.e., with

$$\sum P_N(A_i^*) = \sum P_N(A_i) \tag{42}$$

and

$$\sum P_{SN}(B_i^*) = \sum P_{SN}(B_i) \tag{43}$$

The problem of constructing an optimum sequential test is difficult because the equalities (42) and (43) can be satisfied even when there is no apparent term-by-term relation between the sequences $\{P_N(C_i^*)\}$ and $\{P_N(C_i)\}$. Wald has proposed as optimum the tests in which each of the sequences $\{a_n\}$ and $\{b_n\}$ is constant, that is, $b_1 = b_n$ and $a_1 = a_n$ for all n. Moreover Wald and Wolfowitz[10] proved that these tests are optimum whenever the density functions at successive stages are independent, as can be the case for example when both noise and signal plus noise consist of "random noise." However, this "randomness" is not met with in most applications of the theory of signal detectability at least not in the sense that the hypotheses of Wald and Wolfowitz are satisfied.

Consider a test of fixed length as described in Section 2, with error probabilities F and M. Although the optimum sequential test with these same error probabilities generally requires less time on the average, it has the disadvantage that it will sometimes use much more time than the fixed length test requires. In a conversation with the authors, Professor Mark Kac of Cornell University suggested that the dispersion, or variance, of the sample numbers may be so large as seriously to affect the usefulness of the sequential tests in applications to signal detectability. Certainly this matter should be investigated before a final decision is reached concerning the merits of sequential tests relative to tests on a fixed observation interval. However it is a difficult matter to calculate the variance of the sample numbers. Therefore an electronic simulator is being built at the University of Michigan which will simulate both types of tests and will provide data for ROC curves of both types as well as the distribution of the (sequential) sample numbers.

4. OPTIMUM DETECTION FOR SPECIFIC CASES

4.1 Introduction

The chief conclusion obtained from the general theory of signal detectability presented in Section 2 of this paper is that a receiver which calculates the likelihood ratio for each receiver input is the optimum receiver for detecting signals in noise.

It is the purpose of Section 4 to consider a number of different ensembles of signals with band-limited white Gaussian noise. For each case, a possible receiver design is discussed. The primary emphasis, however, is on obtaining the probability of detection and probability of false alarm, and hence on estimates of optimum receiver performance for the various cases.

The cases which are presented were chosen from the simplest problems in signal detection which closely represent practical situations. They are listed in Table I along with examples of engineering problems in which they find application. In the last two cases the uncertainty in the signal can be varied, and some light is thrown on the relationship between uncertainty and the ability to detect

signals. The variety of examples presented should serve to suggest methods for attacking other simple signal detection problems and to give insight into problems too complicated to allow a direct solution.

The reader will find the discussion of likelihood ratio and its distribution easier to follow if he keeps in mind the connection between a criterion type receiver and likelihood ratio. In an optimum criterion type system, the operator will say that a signal is present whenever the likelihood ratio is above a certain level β. He will say that only noise is present when the likelihood ratio is below β. For each operating level β, there is a false alarm probability and a probability of detection. The false alarm probability is the probability that the likelihood ratio $l(X)$ will be greater than β if no signal is sent; this is by definition the complementary distribution function $F_N(\beta)$. Likewise, the complementary distribution $F_{SN}(\beta)$ is the probability that $l(X)$ will be greater than β if there is signal plus noise, and hence $F_{SN}(\beta)$ is the probability of detection if a signal is sent.

TABLE I

Section	Description of Signal Ensemble	Application
4.4	Signal Known Exactly*	Coherent radar with a target of known range and character
4.5	Signal Known Except for Phase*	Ordinary pulse radar with no integration and with a target of known range and character.
4.6	Signal a Sample of White Gaussian Noise	Detection of noise-like signals; detection of speech sounds in Gaussian noise.
4.7	Detector Output of a Broad Band Receiver	Detecting a pulse of known starting time (such as a pulse from a radar beacon) with a crystal-video or other type broad band receiver.
4.8	A Radar Case (A train of pulses with incoherent phase)	Ordinary pulse radar with integration and with a target of known range and character.
4.10	Signal One of M Orthogonal Signals	Coherent radar where the target is at one of a finite number of non-overlapping positions.
4.11	Signal One of M Orthogonal Signals Known Except for Phase	Ordinary pulse radar with no integration and with a target which may appear at one of a finite number of non-overlapping positions.

4.2 Gaussian Noise

In the remainder of this paper the receiver inputs will be assumed to be defined on a finite observation interval, $0 < t < T$. It will further be assumed that the receiver inputs are series-bandlimited. By the sampling plan C (Section 1.2) any such receiver input $x(t)$ can be reconstructed from sample values of the function taken at points $1/2W$ apart throughout the observation interval, i.e.,

---

* Our treatment of these two fundamental cases is based upon Woodward and Davies' work, but here they are treated in terms of likelihood ratio, and hence apply to criterion type receivers as well as to a posteriori probability type receivers. These first two cases have been solved for the more general problem in which the noise is Gaussian but has an arbitrary spectrum.[11, 12] Those solutions require the use of an infinite sampling plan and are consideredly more involved than the corresponding derivations in this report.

$$x(t) = \sum_{k=1}^{2WT} x_k \psi_k(t), \tag{44}$$

where

$$\psi_k(t) = \frac{\sin \pi 2WT(\frac{t}{T} - \frac{k}{2WT})}{2WT \sin \pi (\frac{t}{T} - \frac{k}{2WT})} \quad \text{and} \quad x_k = x(\tfrac{k}{2W}). \tag{45}$$

Therefore the receiver inputs can be represented by the sample $(x_1, x_2, \ldots, x_{2WT})$. In Section 4 the notation x will be used to denote either the receiver input function x(t) or the sample $(x_1, x_2, \ldots, x_{2WT})$. Similarly the signal $\underset{\sim}{s}(t)$, or simply $\underset{\sim}{s}$, can be represented by the sample $(\underset{\sim}{s}_1, \ldots, \underset{\sim}{s}_{2WT})$, where $\underset{\sim}{s}_k = s(k/2W)$.

Only the probability distributions for receiver inputs x(t) can be specified. The distribution must be given for the receiver inputs both with noise alone and with signal plus noise. The probability distributions are described by giving the probability density functions $f_{SN}(x)$ and $f_N(x)$ for the receiver inputs x.

The probability density function for the receiver inputs with noise alone are assumed to be

$$f_N(x) = \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi N}} \exp \left[- \frac{x_i^2}{2N} \right] \right\}, \tag{46}$$

or

$$f_N(x) = \left(\frac{1}{2\pi N} \right)^{\frac{n}{2}} \exp \left[- \frac{1}{2N} \sum_{i=1}^{n} x_i^2 \right]$$

where n is 2WT and N is the noise power. It can be verified easily that this probability density function is the description of noise which has a Gaussian distribution of amplitude at every time, is stationary, and has the same average power in each of its Fourier components. Thus we shall refer to it as "stationary band-limited white Gaussian noise."

The functions $\psi_k(t)$ are orthogonal and have energy 1/2W, and therefore

$$\sum x_i^2 = 2W \int_0^T \left[x(t) \right]^2 dt \quad , \tag{47}$$

so that

$$f_N(x) = \left(\frac{1}{2\pi N} \right)^{\frac{n}{2}} \exp \left[- \frac{1}{N_o} \int_0^T x(t)^2 \, dt \right] , \tag{48}$$

where $N_o = N/W$ is the noise power per unit bandwidth.

In a practical application, information is given about the signals as they would appear without noise at the receiver input, rather than about the signal plus noise probability density. Then $f_{SN}(x)$ must be calculated from this information and the probability density function $f_N(x)$ for the noise. The noise and the signals will be assumed independent of each other.

If the input to the receiver is the sum of the signal and the noise, then the receiver input x(t) could have been caused by any signal s(t) and noise n(t) = x(t) - s(t). The probability density for the input x in signal plus noise is thus the probability (density) that s(t) and x(t) - s(t) will occur together, averaged over all possible s(t). If the probability of the signals is described by a density function $f_S(s)$, then

$$f_{SN}(x) = \int f_N(x-s) f_S(s)\,ds \tag{49}$$

where the integration is over the entire range of the sample variable s. A more general form is used when the probability of the signals is described by a probability measure P_S; the formula in this case is

$$f_{SN}(x) = \int f_N(x-s)\, dP_S(s). \tag{50}$$

This integral is a Lebesgue integral, and is essentially an "average" of $f_N(x-s)$ over all values of s weighted by the probability P_S. If $f_N(x)$ is taken from Eq. (46), this becomes

$$\begin{aligned} f_{SN}(x) &= \int f_N(x-s)\,dP_S(s) = \left(\frac{1}{2\pi N}\right)^{\frac{n}{2}} \int \exp\left[-\frac{1}{2N}\sum_{i=1}^{n}(x_i-s_i)^2\right] dP_S(s) \\ &= \left(\frac{1}{2\pi N}\right)^{\frac{n}{2}} \exp\left[-\frac{1}{2N}\sum_{i=1}^{n} x_i^2\right] \int \exp\left[-\frac{1}{2N}\sum_{i=1}^{n} s_i^2\right] \exp\left[\frac{1}{N}\sum_{i=1}^{n} x_i s_i\right] dP_S(s) \end{aligned} \tag{51}$$

$$\begin{aligned} f_{SN}(x) &= \int f_N(x-s)\, dP_S(s) = \left(\frac{1}{2\pi N}\right)^{\frac{n}{2}} \int \exp\left[-\frac{1}{N_o}\int_0^T \left[x(t)-s(t)\right]^2 dt\right] dP_S(s) \\ &= \left(\frac{1}{2\pi N}\right)^{\frac{n}{2}} \exp\left[-\frac{1}{N_o}\int_0^T x^2\,dt\right] \int \exp\left[-\frac{1}{N_o}\int_0^T s^2 dt\right] \exp\left[\frac{2}{N_o}\int_0^T xs\,dt\right] dP_S(s) \end{aligned} \tag{52}$$

The factor $\exp\left[-(1/N_o)\int_0^T x^2(t)\,dt\right] = \exp\left[-(1/2N)\sum x_i^2\right]$ can be brought out of the integral since it does not depend on s, the variable of integration. Note that the integral

$$\int_0^T s(t)^2 dt = \frac{1}{2W}\sum s_i^2 = E(s) \tag{53}$$

is the energy* of the expected signal, while

$$\int_0^T x(t)\, s(t)\, dt = \frac{1}{2W}\sum x_i s_i \tag{54}$$

is the cross correlation between the expected signal and the receiver input.

4.3 Likelihood Ratio with Gaussian Noise

Likelihood ratio is defined as the ratio of the probability density functions $f_{SN}(x)$ and $f_N(x)$. With white Gaussian noise it is obtained by dividing Eq. (51) and (52) by (46) and (48) respectively.

---

* This assumes that the circuit impedance is normalized to one ohm.

$$\ell(x) = \int \exp\left[-\frac{E(s)}{N_o}\right] \exp\left[\frac{1}{N}\sum_{i=1}^{n} x_i s_i\right] dP_S(s), \quad \text{or} \tag{55}$$

$$\ell(x) = \int \exp\left[-\frac{E(s)}{N_o}\right] \exp\left[\frac{2}{N_o}\int_0^T x(t)\, s(t)\, dt\right] dP_S(s) \,. \tag{56}$$

If the signal is known exactly or completely specified, the probability for that signal is unity, and the probability for any set of possible signals not containing s is zero. Then the likelihood ratio becomes

$$\ell_s(x) = \exp\left[-\frac{E(s)}{N_o}\right] \exp\left[\frac{1}{N}\sum_{i=1}^{n} x_i s_i\right], \quad \text{or} \tag{57}$$

$$\ell_s(x) = \exp\left[-\frac{E(s)}{N_o}\right] \exp\left[\frac{2}{N_o}\int_0^T x(t)\, s(t)\, dt\right] \,. \tag{58}$$

Thus the general formulas (55) and (56) for likelihood ratio state that $\ell(x)$ is the weighted average of $\ell_s(x)$ over the set of all signals, i.e.,

$$\ell(x) = \int \ell_s(x)\, d\, P_S(s) \,. \tag{59}$$

An equipment which calculates the likelihood ratio $\ell(x)$ for each receiver input x is the optimum receiver. The form of equation (58) suggests one form which this equipment might take. First, for each possible expected signal s, the individual likelihood ratio $\ell_s(x)$ is calculated. Then these numbers are averaged. Since the set of expected signals is often infinite, this direct method is usually impractical. It is frequently possible in particular cases to obtain by mathematical operations on equation (58) a different form for $\ell(x)$ which can be recognized as the response of a realizable electronic equipment, simpler than the equipment specified by the direct method. It is essentially this which is done in the following paragraphs.

If the distribution function $P_S(s)$ depends on various parameters such as carrier phase, signal energy, or carrier frequency, and if the distributions in these parameters are independent, the expression for likelihood ratio can be simplified somewhat. If these parameters are indicated by r_1, r_2, ... , r_n, and the associated probability density functions are denoted by $f_1(r_1)$, $f_2(r_2)$, ... , $f_n(r_n)$, then

$$d\, P_S(s) = f_1(r_1) \cdots f_n(r_n)\, dr_1 \cdots dr_n \,.$$

The likelihood ratio becomes

$$\begin{aligned} \ell(x) &= \int \cdots \int \ell_s(x)\, f_1(r_1) \cdots f_n(r_n)\, dr_1 \cdots dr_n \\ &= \int \left[f_n(r_n) \cdots \left[\int f_1(r_1)\, \ell_s(x)\, dr_1 \right] \cdots \right] dr_n \,. \end{aligned} \tag{60}$$

Thus the likelihood ratio can be found by averaging $\ell_s(x)$ with respect to the parameters.

4.4 The Case of a Signal Known Exactly

The likelihood ratio for the case when the signal is known exactly has already been presented in Section 4.3.

$$\ell(x) = \exp\left[-\frac{E}{N_o}\right] \exp\left[\frac{1}{N}\sum_{i=1}^{n} x_i s_i\right] , \tag{61}$$

$$\ell(x) = \exp\left[-\frac{E}{N_o}\right] \exp\left[\frac{2}{N_o}\int_0^T x(t)\, s(t)\, dt\right] \tag{62}$$

As the first step in finding the distribution functions for $\ell(x)$, it is convenient to find the distribution for $(1/N)\sum x_i s_i$ when there is noise alone. Then the input $x = (x_1, x_2, \ldots, x_n)$ is due to white Gaussian noise. It can be seen from Eq. (46) that each x_i has a normal distribution with zero mean and variance $N = WN_o$ and that the x_i are independent. Because the s_i are constants depending on the signal to be detected, $s = (s_1, s_2, \ldots, s_n)$, each summand $(x_i s_i)/N$ has a normal distribution with mean s_i/N times the mean of x_i, and with variance $(s_i/N)^2$ times the variance of x_i, which are zero and s_i^2/N respectively. Because the x_i are independent, the summands $(s_i x_i)/N$ are independent, each with normal distribution, and therefore their sum has a normal distribution with mean the sum of the means -- i.e., zero -- and variance the sum of the variances.

$$\sum \frac{s_i^2}{N} = \frac{2WE(s)}{N} = \frac{2E}{N_o} = 2 \times \frac{\text{Signal Energy}}{\text{Noise Power Per Unit Bandwidth}} . \tag{63}$$

The distribution for $(1/N)\sum x_i s_i$ with noise alone is thus normal with zero mean and variance $2E/N_o$. Recalling from Eq. (61)

$$\ell(x) = \exp\left[-\frac{E}{N_o} + \frac{1}{N}\sum x_i s_i\right] , \tag{64}$$

one sees that the distribution for $(1/N)\sum x_i s_i$ can be used directly by introducing α defined by

$$\beta = \exp\left[-\frac{E}{N_o} + \alpha\right] , \quad \text{or } \alpha = \frac{E}{N_o} + \ell n\, \beta . \tag{65}$$

The inequality $\ell(x) \geq \beta$ is equivalent to $(1/N)\sum x_i s_i \geq \alpha$, and therefore

$$F_N(\beta) = \sqrt{\frac{N_o}{4\pi E}} \int_\alpha^\infty \exp\left[-\frac{1}{2}\frac{N_o}{2E} y^2\right] dy . \tag{66}$$

The distribution for the case of signal plus noise can be found by using Eq. (19), which states that

$$\left(\frac{d\, P_{SN}(A(\beta))}{d\, P_N(A(\beta))}\right)_{\text{at } \beta=\beta_o} = \beta_o . \tag{67}$$

Because these probabilities are equal to the complimentary distribution functions for likelihood ratio, this can be written as

$$d\, F_{SN}(\beta) = \beta\, d\, F_N(\beta) . \tag{68}$$

Differentiating Eq. (66),

$$dF_N(\beta) = -\sqrt{\frac{N_o}{4\pi E}} \exp\left(-\frac{N_o \alpha^2}{4E}\right) d\alpha , \tag{69}$$

and combining (65), (68), and (69), one obtains

$$dF_{SN}(\beta) = -\sqrt{\frac{N_o}{4\pi E}} \exp\left[-\frac{E}{N_o} + \alpha - \frac{N_o\alpha^2}{4E}\right] d\alpha . \tag{70}$$

Thus

$$F_{SN}(\beta) = \sqrt{\frac{N_o}{4\pi E}} \int_{\alpha}^{\infty} \exp\left[-\frac{N_o}{4E}\left(y - \frac{2E}{N_o}\right)^2\right] dy . \tag{71}$$

In summary, α and therefore $\ell n\ \beta$, have normal distributions with signal plus noise as well as with noise alone; the variance of each distribution is $2E/N_o$, and the difference of the means is $2E/N_o$.

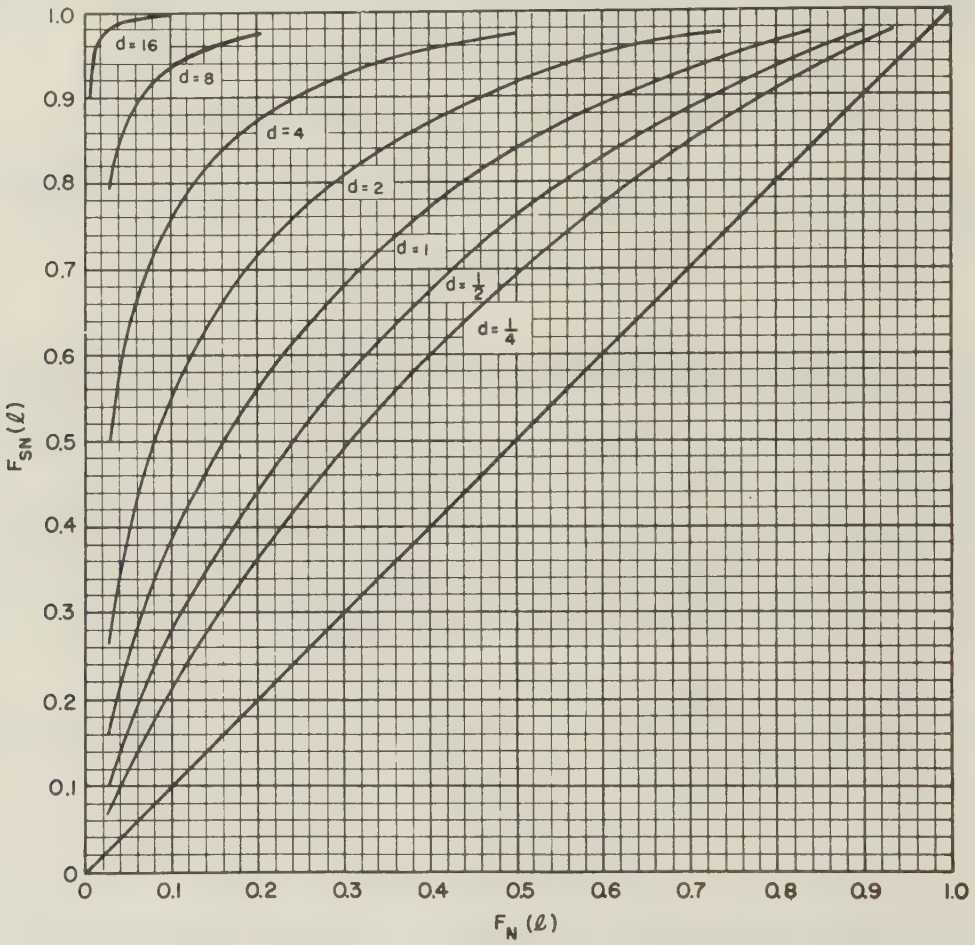


FIG. 2

RECEIVER OPERATING CHARACTERISTIC

$\ell n\ \ell$ IS A NORMAL DEVIATE WITH $\sigma_N^2 = \sigma_{SN}^2$, $(M_{SN}-M_N)^2 = d\cdot\sigma_N^2$

The receiver operating characteristic curves in Figs. 2 and 3* are plotted for any case in which

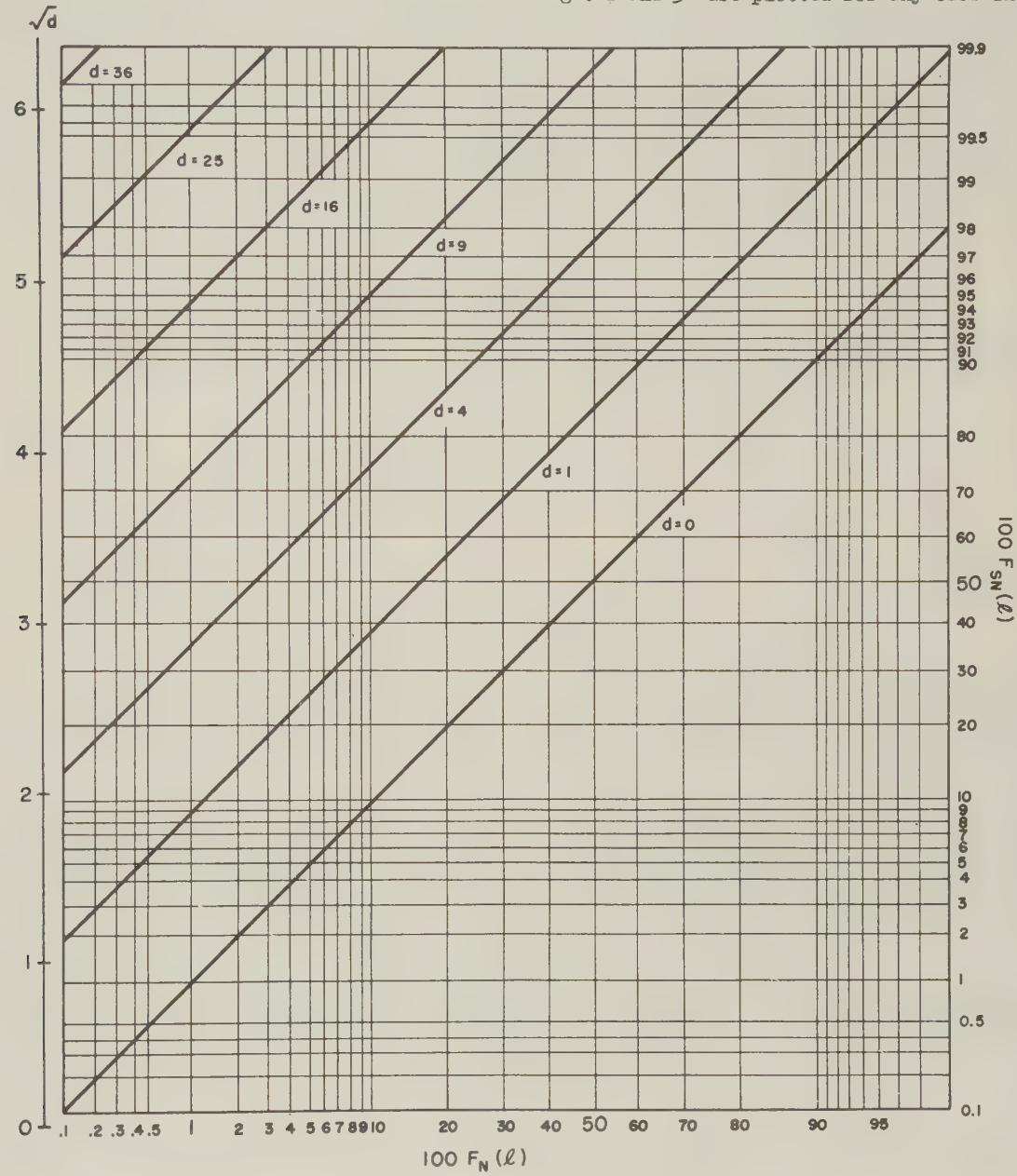


FIG. 3
RECEIVER OPERATING CHARACTERISTIC.

$\ell n\ \ell$ IS A NORMAL DEVIATE, $\sigma_{SN}^2 = \sigma_N^2$, $(M_{SN} - M_N)^2 = d\ \sigma_N^2$

---

* In Fig. 3, the receiver operating characteristic curves are plotted on "double probability" paper. On this paper both axes are linear in the error function erf $(x) = (1/\sqrt{2\pi}) \cdot \int_{-\infty}^{x} \exp\left[-t^2/2\right] dt$; this makes the receiver operating characteristic straight lines.

$\ell n\ \ell$ has a normal distribution with the same variance both with noise alone and with signal plus noise. The parameter d in this figure is equal to the square of the difference of the means, divided by the variance. These receiver operating characteristic curves apply to the case of the signal known exactly, with $d = 2E/N_o$.

Eq. (62) describes what the ideal receiver should do for this case. The essential operation in the receiver is obtaining the correlation, $\int_0^T s(t)x(t)dt$. The other operations, multiplying by a constant, adding a constant, and taking the exponential function, can be taken care of simply in the calibration of the receiver output. Electronic means of obtaining cross correlation have been developed recently.[13]

If the form of the signal is simple, there is a simple way to obtain this cross correlation.[6,7] Suppose h(t) is the impulse response of a filter. The response $e_o(t)$ of the filter to a voltage x(t) is

$$e_o(t) = \int_{-\infty}^{t} x(\tau)\, h(t-\tau)\, d\tau \ . \tag{72}$$

If a filter can be synthesized so that

$$\begin{aligned} h(t) &= s(T-t) && 0 \leqq t \leqq T \\ h(t) &= 0 && \text{otherwise,} \end{aligned} \tag{73}$$

then

$$e_o(T) = \int_0^T x(\tau)\, s(\tau)\, d\tau \quad , \tag{74}$$

so that the response of this filter at time T is the cross correlation required. Thus, the ideal receiver consists simply of a filter and amplifiers.

It should be noted that this filter is the same, except for a constant factor, as that specified when one asks for the filter which maximizes peak signal to average noise power ratio.[14]

4.5 Signal Known Except for Carrier Phase

The signal ensemble considered in this section consists of all signals which differ from a given amplitude and frequency modulated signal only in their carrier phase, and all carrier phases are assumed equally likely.

$$s(t) = f(t) \cos\left(\omega t + \phi(t) - \theta\right) \ . \tag{75}$$

Since the unknown phase angle θ has a uniform distribution,

$$dP_S(\theta) = \frac{1}{2\pi}\, d\theta. \tag{76}$$

The likelihood ratio can be found by applying Eq.(56), and since the signal energy E(s) is the same for all values of the carrier phase θ,

$$\ell(x) = \exp\left[-\frac{E}{N_o}\right] \int \exp\left[\frac{1}{N}\sum x_i s_i\right] dP_S(s) \quad . \tag{77}$$

Expanding s into the coefficients of $\cos\theta$ and $\sin\theta$ will be helpful:

$$s(t) = f(t)\cos\left(\omega t+\phi(t)\right)\cos\theta + f(t)\sin\left(\omega t+\phi(t)\right)\sin\theta \quad , \tag{78}$$

and

$$\frac{1}{N}\sum x_i s_i = \cos\theta \frac{1}{N}\sum x_i f(t_i)\cos\left(\omega t_i + \phi(t_i)\right) + \sin\theta \frac{1}{N}\sum x_i f(t_i)\sin\left(\omega t_i + \phi(t_i)\right) \quad * \tag{79}$$

Because we wish to integrate with respect to θ to find the likelihood ratio, it is easiest to introduce parameters similar to polar coordinates (r, θ_o) such that

$$\frac{1}{N} r\cos\theta_o = \frac{1}{N}\sum x_i f(t_i)\cos\left(\omega t_i + \phi(t_i)\right)$$
$$\frac{1}{N} r\sin\theta_o = \frac{1}{N}\sum x_i f(t_i)\sin\left(\omega t_i + \phi(t_i)\right) \tag{80}$$

and therefore

$$\frac{1}{N}\sum x_i s_i = \frac{r}{N}\cos(\theta - \theta_o) \quad . \tag{81}$$

Using this form the likelihood ratio becomes

$$\ell(x) = \exp\left[-\frac{E}{N_o}\right]\int_0^{2\pi} \exp\left[\frac{r}{N}\cos(\theta - \theta_o)\right]\frac{d\theta}{2\pi}$$
$$= \exp\left[-\frac{E}{N_o}\right] I_o\left(\frac{r}{N}\right) \tag{82}$$

where I_o is the Bessel function of zero order and pure imaginary argument.

I_o is a strictly monotone increasing function, and therefore the likelihood ratio will be greater than a value β if and only if r/N is greater than some value corresponding to β.

In the previous section it was shown that the sum $(1/N)\sum x_i s_i$ has a normal distribution with zero mean and variance $2E/N_o$ if the receiver input $x(t)$ is due to noise alone; E is the energy of the signal known exactly, $s(t)$, and N_o is the noise power per cycle. Since $f(t)\cos(\omega t + \phi(t))$ and $f(t)\sin(\omega t + \phi(t))$ are signals known exactly, both $(r/N)\cos\phi_o$ and $(r/N)\sin\phi_o$ have normal distributions with zero mean and variance $2E/N_o$. The probability that due to noise alone $r/N = \sqrt{(r/N\cos\theta_o)^2 + (r/N\sin\theta_o)^2}$ will exceed any fixed value, is given by the well known chi-square distribution for two degrees of freedom, $K_2(\alpha^2)$. The proper normalization yielding zero mean and unit variance requires that the variable be $(r/N)\sqrt{N_o/2E(s)}$, that is

$$P_N\left(\frac{r}{N}\sqrt{\frac{N_o}{2E}} \geq \alpha\right) = K_2(\alpha^2) = \exp\left[-\frac{\alpha^2}{2}\right]. \quad ** \tag{83}$$

---

* t_i denotes the i^{th} sampling time, i.e., $t_i = i/2W$.

** The symbol $P(x \geq \alpha)$ denotes the probability that the variable x is not less than the constant α.

If α is defined by the equation

$$\beta = \exp\left[-\frac{E}{N_o}\right] I_o\left(\sqrt{\frac{2E}{N_o}}\ \alpha\right), \tag{84}$$

the distribution for $\ell(x)$ in the presence of noise alone is in the simple form

$$F_N(\beta) = \exp\left[-\frac{\alpha^2}{2}\right] . \tag{85}$$

It follows from (85) that

$$dF_N(\beta) = -\alpha \exp\left[-\frac{\alpha^2}{2}\right] d\alpha . \tag{86}$$

If in equation (68), namely

$$\beta\ dF_N(\beta) = dF_{SN}(\beta) , \tag{87}$$

β is replaced by the expression given in (84) and $dF_N(\beta)$ is replaced by that given in (86), then

$$dF_{SN}(\beta) = -\exp\left[-\frac{E}{N_o}\right]\alpha \exp\left[-\frac{\alpha^2}{2}\right] I_o\left(\sqrt{\frac{2E}{N_o}}\ \alpha\right) d\alpha \tag{88}$$

is obtained. Integration of (88) yields

$$F_{SN}(\beta) = \exp\left[-\frac{E}{N_o}\right] \int_{\alpha}^{\infty} \alpha \exp\left[-\frac{\alpha^2}{2}\right] I_o\left(\sqrt{\frac{2E}{N_o}}\ \alpha\right) d\alpha . \tag{89}$$

Eqs. (85) and (89) yield the receiver operating characteristic in parametric form, and Eq. (84) gives the associated operating levels.[15] These are graphed in Fig. 4 for some of the same values of signal energy to noise power per unit bandwidth as were used when the phase angle was known exactly, Figs. 2 and 3, so that the effect of knowing the phase can be easily seen.

If the signal is sufficiently simple so that a filter could be synthesized to match the expected signal for a given carrier phase Θ as in the case of a signal known exactly, then there is a simple way to design a receiver to obtain likelihood ratio. For simplicity let us consider only amplitude modulated signals ($\phi(t)=0$) in Eq. (75). Let us also choose $\Theta = 0$. (Any phase could have been chosen.) Then the filter has impulse response

$$\begin{aligned} h(t) &= f(T-t)\cos\left[\omega(T-t)\right] && 0 \le t \le T \\ &= 0 && \text{otherwise.} \end{aligned} \tag{90}$$

The output of the filter in response to x(t) is then

$$\begin{aligned} e_o(t) &= \int_{-\infty}^{t} x(\tau)\, h(t-\tau)\, d\tau = \int_{t-T}^{t} x(\tau)\, f(\tau+T-t)\cos\omega(\tau+T-t)\, d\tau \\ &= \cos\omega(T-t) \int_{t-T}^{t} x(\tau)\, f(\tau+T-t)\cos\omega\tau\ d\tau \\ &\quad - \sin\omega(T-t) \int_{t-T}^{t} x(\tau)\, f(\tau+T-t)\sin\omega\tau\ d\tau . \end{aligned} \tag{91}$$

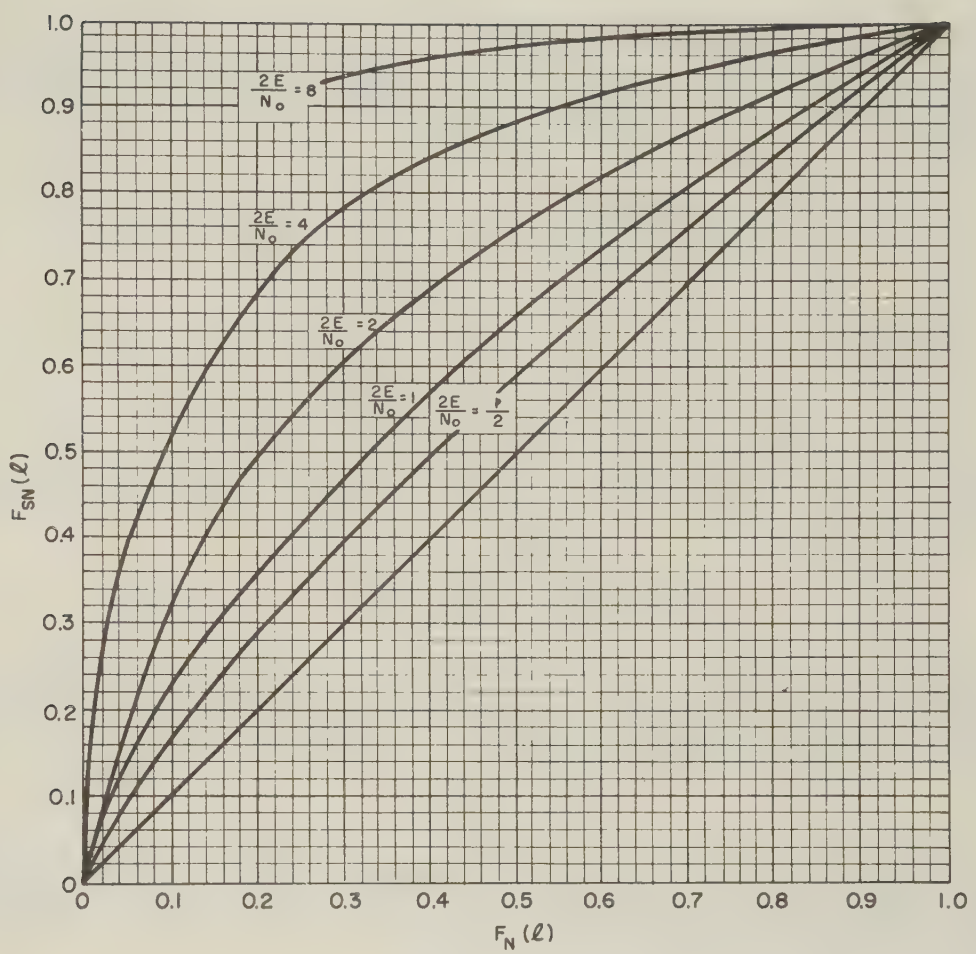


Fig. 4

RECEIVER OPERATING CHARACTERISTIC.

SIGNAL KNOWN EXCEPT FOR PHASE.

The envelope of the filter output will be the square root of the sum of the squares of the integrals;* and the envelope at time T will be proportional to r/N, since

$$\left(\frac{r}{2W}\right)^2 = \left[\int_0^T x(\tau)\, f(\tau) \cos \omega\tau \, d\tau\right]^2 + \left[\int_0^T x(\tau)\, f(\tau) \sin \omega\tau \, d\tau\right]^2, \qquad (92)$$

which can be identified as the square of the envelope of $e_o(t)$ at time T. If the input x(t) passes through the filter with an impulse response given by Eq. (90), then through a linear detector, the output will be $(N_o/2)r/N$ at time T. Because the likelihood ratio, Eq. (82), is a known monotone function of r/N, the output can be calibrated to read the likelihood ratio of the input.

---

* If the line spectrum of s(t) is zero at zero frequency and at all frequencies equal to or greater than $2W/2\pi$, then it can be shown that these integrals contain no frequencies as high as $W/2\pi$.

4.6 Signal Consisting of a Sample of White Gaussian Noise

Suppose the values of the signal voltage at the sample points are independent Gaussian random variables with zero mean and variance S, the signal power. The probability density due to signal plus noise is also Gaussian, since signal plus noise is the sum of two Gaussian random variables:

$$f_{SN}(x) = \left(\frac{1}{2\pi(N+S)}\right)^{\frac{n}{2}} \exp\left[-\frac{1}{2}\frac{1}{N+S}\sum x_i^2\right], \tag{93}$$

where n = 2WT.

The likelihood ratio is

$$\ell(x) = \left(\frac{N}{N+S}\right)^{\frac{n}{2}} \exp\left[\frac{1}{2}\frac{1}{N}\sum x_i^2 - \frac{1}{2}\frac{1}{N+S}\sum x_i^2\right] . \tag{94}$$

In determining the distribution functions for ℓ, it is convenient to introduce the parameter α, defined by the equation

$$\beta = \left(\frac{N}{N+S}\right)^{\frac{n}{2}} \exp\left(\frac{S}{N+S}\frac{\alpha^2}{2}\right) . \tag{95}$$

Then the condition $\ell(x) \geq \beta$ is equivalent to the condition that $(1/N)\sum x_i^2 \geq a^2$. In the presence of noise alone the random variables $x_i/\sqrt{N}$ have zero mean and unit variance, and they are independent. Therefore, the probability that the sum of the squares of these variables will exceed a^2 is the chi-square distribution with n degrees of freedom, i.e.,

$$F_N(\beta) = K_n(\alpha^2) . \tag{96}$$

Similarly, in the presence of signal plus noise the random variables $x_i/\sqrt{N+S}$ have zero mean and unit variance. The condition $(1/N)\sum x_i^2 \geq a^2$ is the same as requiring that $(1/N+S)\sum x_i^2 \geq (N/N+S)a^2$, and again making use of the chi-square distribution,

$$F_{SN}(\beta) = K_n\left(\frac{N}{N+S}\alpha^2\right) . \tag{97}$$

For large values of n, the chi-square distribution is approximately normal over the center portion; more precisely,[16] for $\alpha^2 >> 0$,

$$F_N(\beta) = K_n(\alpha^2) \approx \frac{1}{\sqrt{2\pi}} \int_{\sqrt{2\alpha^2} - \sqrt{2n-1}}^{\infty} \exp\left[-\frac{1}{2}y^2\right] dy \tag{98}$$

and

$$F_{SN}(\beta) = K_n\left(\frac{N}{N+S}\alpha^2\right) \approx \frac{1}{\sqrt{2\pi}} \int_{\sqrt{\frac{2N\alpha^2}{N+S}} - \sqrt{2n-1}}^{\infty} \exp\left[-\frac{1}{2}y^2\right] dy . \tag{99}$$

If the signal energy is small compared to that of the noise, $\sqrt{N/N+S}$ is nearly unity and both distribu-

tions have nearly the same variance. Then Figs. 2 and 3 apply to this case too, with the value of d given by

$$d = (2n-1)\left(1 - \sqrt{\frac{N}{N+S}}\right)^2 \tag{100}$$

For these small signal to noise ratios and large samples, there is a simple relation between signal to noise ratio, the number of samples, and the detection index d.

$$1 - \sqrt{\frac{N}{N+S}} \approx \frac{1}{2}\frac{S}{N} \qquad \text{for } \frac{S}{N} << 1 \quad , \quad \text{and}$$
$$d \approx \frac{nS^2}{2N^2} \tag{101}$$

Two signal to noise ratios, $(S/N)_1$ and $(S/N)_2$, will give approximately the same operating characteristic if the corresponding numbers of sample points, n_1 and n_2, satisfy

$$\frac{n_1}{n_2} = \frac{\left(\frac{S}{N}\right)_1^2}{\left(\frac{S}{N}\right)_2^2} \tag{102}$$

By Eq. (94), the likelihood is a monotone function of $\sum x_i^2$. But the output of an energy detector,

$$e_o(t) = \int_0^T \left[x(t)\right]^2 dt = \frac{1}{2W}\sum x_i^2 \tag{103}$$

is proportional to $\sum x_i^2$. Therefore an energy detector can be calibrated to read likelihood ratio, and hence can be used as an optimum receiver in this case.

4.7 Video Design of a Broad Band Receiver

The problem considered in this section is represented schematically in Fig. 5. The signals

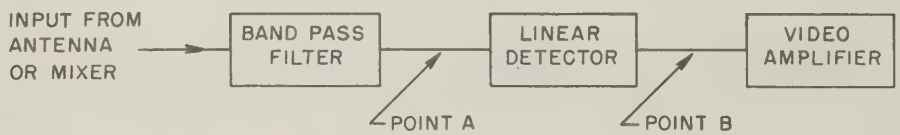


Fig. 5

BLOCK DIAGRAM OF A BROAD BAND RECEIVER

and noise are assumed to have passed through a band pass filter, and at the output of the filter, point A on the diagram, they are assumed to be limited in spectrum to a band of width W and center frequency $\omega/2\pi > W/2$. The noise is assumed to be Gaussian noise with a uniform spectrum over the band. The signals and noise then pass through a linear detector. The output of the detector is the envelope of the signals and noise as they appeared at point A; all knowledge of the phase of the receiver input is lost at point B. The signals and noise as they appear at point B are considered receiver inputs,

and the theory of signal detectability is applied to these video inputs to ascertain the best video design and the performance of such a system. The mathematical description of the signals and noise will be given for the signals and noise as they appear at point A. The envelope functions, which appear at point B, will be derived, and the likelihood ratio and its distribution will be found for these envelope functions.

The only case which will be considered here is the case in which the amplitude of the signal as it would appear at point A is a known function of time.

Any function at point A will be band limited to a band of width W and center frequency $\omega/2\pi > W/2$. Any such function f(t) can be expanded as follows:

$$f(t) = x(t) \cos \omega t + y(t) \sin \omega t \tag{105}$$

where x(t) and y(t) are band limited to frequencies no higher than W/2, and hence can themselves* be expanded by sampling plan C, yielding

$$f(t) = \sum_i \left[x\left(\frac{i}{W}\right) \psi_i(t) \cos \omega t + y\left(\frac{i}{W}\right) \psi_i(t) \sin \omega t \right]. \tag{106}$$

The amplitude of the function f(t) is

$$r(t) = \sqrt{\left[x(t)\right]^2 + \left[y(t)\right]^2} \tag{107}$$

and thus the amplitude at the i^{th} sampling point is

$$r\left(\frac{i}{W}\right) = r_i = \sqrt{x_i^2 + y_i^2} \quad . \tag{108}$$

The angle

$$\theta_i = \arctan \frac{y_i}{x_i} = \arccos \frac{x_i}{r_i} \tag{109}$$

might be considered the phase of f(t) at the i^{th} sampling point. The function f(t) then might be described by giving the r_i and θ_i rather than the x_i and y_i.

Let us denote by x_i, y_i, or r_i, θ_i, the sample values for a receiver input after the filter (i.e., at the point A in Fig. 5). Let a_i, b_i, or f_i, ϕ_i, denote the sample values for the signal as it would appear at point A if there were no noise. The envelope of the signal, hence the amplitude sample values f_i, are assumed known. Let us denote by $F_S(\phi_1, \phi_2, \ldots, \phi_{n/2})$ the distribution function of the phase sample values ϕ_i. The probability density function for the input at A when there is white Gaussian noise and no signal, with $n = 2WT$, is

$$f_N(x, y) = \left(\frac{1}{2\pi N}\right)^{\frac{n}{2}} \exp \left[- \frac{1}{2N} \sum_{i=1}^{n/2} x_i^2 + \sum_{i=1}^{n/2} y_i^2 \right] \tag{110}$$

and for signal plus noise, it is

$$f_{SN}(x, y) = \left(\frac{1}{2\pi N}\right)^{\frac{n}{2}} \int_R \exp \left[- \frac{1}{2N} \left(\sum_{i=1}^{n/2} (x_i - a_i)^2 + \sum_{i=1}^{n/2} (y_i - b_i)^2 \right) \right] dP_S(a_i b_i) \quad . \tag{111}$$

---

* Because any function f(t) at A has no frequency greater than $(\omega/2\pi) + (W/2)$, the usual sampling plan C might have been used on f(t). However, the distribution in noise alone, $f_N(x_i)$, would probably not be applicable.

Expressed in terms of the (r, θ) sample values, Eq. (110) and Eq. (111) become

$$f_N(r,\theta) = \left(\frac{1}{2\pi N}\right)^{\frac{n}{2}} \prod_{i=1}^{n/2} r_i \exp\left[-\frac{1}{2N}\sum_{i=1}^{n/2} r_i^2\right], \tag{112}$$

and

$$f_{SN}(r,\theta) = \left(\frac{1}{2\pi N}\right)^{\frac{n}{2}} \prod_{i=1}^{n/2} r_i \int_R \exp\left[-\frac{1}{2N}\sum_{i=1}^{n/2}\left\{r_i^2+f_i^2-2r_i f_i \cos(\theta_i-\phi_i)\right\}\right] dF_S\left(\phi_1, \ldots, \phi_{\frac{n}{2}}\right). \tag{113}$$

The factors $\prod r_i$ are introduced because they are the Jacobian of the transformation from the x, y sampling plan to the r, θ sampling plan.[16, *]

The probability density function for r alone, i.e., the density function for the output of the detector, is obtained simply by integrating the density functions for r and θ with respect to θ.

$$f_N(r) = \int_0^{2\pi}\int_0^{2\pi}\cdots\int_0^{2\pi} f_N(r_i,\theta_i)\, d\theta_1\, d\theta_2 \cdots d\theta_{\frac{n}{2}}, \tag{114}$$

or

$$f_N(r) = \left(\frac{1}{N}\right)^{\frac{n}{2}} \prod_{i=1}^{n/2} r_i \exp\left[-\frac{1}{2N}\sum_{i=1}^{n/2} r_i^2\right],$$

and

$$f_{SN}(r) = \int_0^{2\pi}\int_0^{2\pi}\cdots\int_0^{2\pi} f_{SN}(r_i,\theta_i)\, d\theta_1\, d\theta_2\cdots d\theta_{\frac{n}{2}},$$

or

$$f_{SN}(r) = \left(\frac{1}{N}\right)^{\frac{n}{2}} \int_R \prod_{i=1}^{n/2} r_i \exp\left[-\frac{1}{2N}\sum_{i=1}^{n/2}(r_i^2+f_i^2)\right] \prod_{i=1}^{n/2} I_o\left(\frac{r_i f_i}{N}\right) dF(\phi_1,\phi_2\cdots\phi_{\frac{n}{2}}), \tag{115}$$

or

$$f_{SN}(r) = \left(\frac{1}{N}\right)^{\frac{n}{2}} \prod_{i=1}^{n/2} r_i\, I_o\left(\frac{r_i f_i}{N}\right) \exp\left[-\frac{1}{2N}\sum_{i=1}^{n/2}(r_i^2+f_i^2)\right].$$

Notice that the probability density for r is completely independent of the distribution which the ϕ_i had; all information about the phase of the signals has been lost.

The likelihood ratio for a video input r(t), is

$$\ell(r) = \frac{f_{SN}(r)}{f_N(r)} = \exp\left[-\frac{1}{2N}\sum_{i=1}^{n/2} f_i^2\right] \prod_{i=1}^{n/2} I_o\left(\frac{r_i f_i}{N}\right). \tag{116}$$

---

* For example, in two dimensions, $f_N(x,y)dx\,dy = f_N(r,\theta)\, r\, dr\, d\theta$.

Again it is more convenient to work with the logarithm of the likelihood ratio. Thus

$$\frac{1}{2N}\sum_{i=1}^{n/2} f_i^2 = \frac{W}{2N}\int \left[f(t)\right]^2 dt = \frac{E}{N_o}, \text{ and} \tag{117}$$

$$\ell n\ \ell(r) = -\frac{E}{N_o} + \sum_{i=1}^{n/2} \ell n\ I_o\left(\frac{r_i f_i}{N}\right), \tag{118}$$

which is approximately

$$\ell n\ \ell\big(r(t)\big) = -\frac{E}{N_o} + W\int_0^T \ell n\ I_o\left(\frac{r(t)\ f(t)}{N}\right)dt\ . \tag{119}$$

The function $\ell n\ I_o(x)$ is approximately the parabola $x^2/4$ for small values of x and is nearly linear for large values of x. Thus, the expression for likelihood ratio might be approximated by

$$\ell n\,\ell\big(r(t)\big) = -\frac{E}{N_o} + \frac{W}{4N^2}\int_0^T \left[r(t)\right]^2\left[f(t)\right]^2 dt \tag{120}$$

for small signals, and by

$$\ell n\,\ell\big(r(t)\big) = C_1 + C_2\int_0^T r(t)\ f(t)\ dt \tag{121}$$

for large signals, where C_1 and C_2 are chosen to approximate $\ell n\ I_o$ best in the desired range.

The integrals in Eqs. (120) and (121) can be interpreted as cross correlations. Thus the optimum receiver for weak signals is a square law detector, followed by a correlator which finds the cross correlation between the detector output and $(f(t))^2$, the square of the envelope of the expected signal. For the case of large signal to noise ratio, the optimum receiver is a linear detector, followed by a correlator which has for its output the cross correlation of the detector output and f(t), the amplitude of the expected signal.

The distribution function for $\ell(r)$ cannot be found easily in this case. The approximation developed here will apply to the receiver designed for low signal to noise ratio, since this is the case of most interest in detection studies. An analogous approximation for the large signal to noise ratios would be even easier to derive.

First we shall find the mean and standard deviation for the distribution of the logarithm of the likelihood ratio as shown above,

$$\ell n\ \ell(r) \approx -\frac{1}{2N}\sum f_i^2 + \frac{1}{4N^2}\sum_{i=1}^{n/2} r_i^2 f_i^2\ , \tag{122}$$

for the case of small signal to noise ratio. The probability density functions for each r_i are

$$g_{SN}(r_i) = \frac{r_i}{N}\ \exp\left[-\frac{r_i^2+f_i^2}{2N}\right] I_o\left[\frac{r_i f_i}{N}\right], \text{ and} \tag{123}$$

$$g_N(r_i) = \frac{r_i}{N}\ \exp\left[-\frac{r_i^2}{2N}\right]\ .$$

The notation $g_N(r_i)$ and $g_{SN}(r_i)$ is used to distinguish these from the joint distributions of all the r_i which were previously called $f_N(r)$ and $f_{SN}(r)$. The mean of each term $r_i^2 f_i^2/4N^2$ in the sum in Eq. (122) is

$$\mu_{SN}\left(\frac{r_i^2 f_i^2}{4N^2}\right) = \frac{f_i^2}{4N}\int_0^\infty \frac{r_i^2}{N}\ g_{SN}(r_i)\ dr_i\ , \text{ or} \tag{124}$$

$$\mu_{SN}\left(\frac{r_i^2 f_i^2}{4N^2}\right) = \frac{f_i^2}{4N}\int_0^\infty \frac{r_i^3}{N^2}\exp\left[-\frac{(r_i^2+f_i^2)}{2N}\right] I_o\left(\frac{r_i f_i}{N}\right)\, dr_i \ .$$

Similarly,

$$\mu_N\left(\frac{r_i^2 f_i^2}{4N^2}\right) = \frac{f_i^2}{4N}\int_0^\infty \frac{f_i^2}{N}\, g_N(r_i)\, dr_i = \frac{f_i^2}{4N}\int_0^\infty \frac{r_i^3}{N^2}\exp\left[-\frac{r_i^2}{2N}\right] dr_i \qquad (124)$$

The second moment of each term $r_i^2 f_i^2/4N^2$ is

$$\mu_{SN}\left(\frac{r_i^4 f_i^4}{16N^4}\right) = \frac{f_i^4}{16N^2}\int_0^\infty \frac{r_i^4}{N^2}\, g_{SN}(r_i)\, dr_i \ , \text{ or}$$

$$\mu_{SN}\left(\frac{r_i^4 f_i^4}{16N^4}\right) = \frac{f_i^4}{16N^2}\int_0^\infty \frac{r_i^5}{N^3}\exp\left[-\frac{(r_i^2+f_i^2)}{2N}\right] I_o\left(\frac{r_i f_i}{N}\right) dr_i \ . \qquad (125)$$

Similarly,

$$\mu_N\left(\frac{r_i^4 f_i^4}{16N^4}\right) = \frac{f_i^4}{16N^2}\int_0^\infty \frac{r_i^4}{N^2}\, g_N(r_i)\, dr_i \ , \text{ or}$$

$$\mu_N\left(\frac{r_i^4 f_i^4}{16N^4}\right) = \frac{f_i^4}{16N^2}\int_0^\infty \frac{r_i^5}{N^3}\exp\left[-\frac{r_i^2}{2N}\right] dr_i \ .$$

The integrals for the case of noise alone can be evaluated easily:

$$\mu_N\left(\frac{r_i^2 f_i^2}{4N^2}\right) = \frac{f_i^2}{2N} \ , \qquad (126)$$

and

$$\mu_N\left(\frac{r_i^4 f_i^4}{16N^4}\right) = \frac{f_i^4}{2N^2} \ .$$

The integrals for the case of signal plus noise can be evaluated in terms of the confluent hypergeometric function, which turns out for the cases above to reduce to a simple polynomial. The required formulas are collected in convenient form in Threshold Signals[5] on page 174. The results are

$$\mu_{SN}\left(\frac{r_i^2 f_i^2}{4N^2}\right) = \frac{1}{2}\,\frac{f_i^2}{N}\left(1+\frac{f_i^2}{2N}\right) \ ,$$

and

$$\mu_{SN}\left(\frac{r_i^4 f_i^4}{16N^4}\right) = \frac{1}{2}\,\frac{f_i^4}{N^2}\left(1+\frac{f_i^2}{N}+\frac{f_i^4}{8N^2}\right) \ . \qquad (127)$$

Since

$$\sigma^2(Z) = \mu(Z^2) - \left[\mu(Z)\right]^2, \tag{128}$$

the variances of $r_i^2 f_i^2/4N^2$ are

$$\sigma^2_{SN}\left(\frac{r_i^2 f_i^2}{4N^2}\right) = \frac{1}{4}\,\frac{f_i^4}{N^2}\left(1 + \frac{f_i^2}{N}\right)$$

and

$$\sigma^2_{N}\left(\frac{r_i^2 f_i^2}{4N^2}\right) = \frac{f_i^4}{4N^2}\ . \tag{129}$$

For the sum of independent random variables, the mean is the sum of the means of the terms and the variance is the sum of the variances. Therefore the means of $\ell n\, \ell(r)$ are

$$\mu_{SN}\left(\ell n\, \ell(r)\right) = -\frac{1}{2N}\sum_{i=1}^{n/2} f_i^2 + \sum_{i=1}^{n/2}\left[\frac{1}{2}\,\frac{f_i^2}{N} + \frac{1}{4}\,\frac{f_i^4}{N^2}\right] = \sum_{i=1}^{n/2}\frac{f_i^4}{4N^2}$$

and

$$\mu_{N}\left(\ell n\, \ell(r)\right) = -\sum_{i=1}^{n/2}\frac{f_i^2}{2N} + \frac{1}{2}\sum_{i=1}^{n/2}\frac{f_i^2}{N} = 0\ , \tag{130}$$

and the variances of $\ell n\, \ell(r)$ are

$$\sigma^2_{SN}\left(\ell n\, \ell(r)\right) = \sum_{i=1}^{n/2}\left(\frac{1}{4}\,\frac{f_i^4}{N^2} + \frac{1}{4}\,\frac{f_i^6}{N^3}\right)$$

and

$$\sigma^2_{N}\left(\ell n\, \ell(r)\right) = \sum_{i=1}^{n/2}\frac{f_i^4}{4N^2}\ . \tag{131}$$

If the distribution functions of $\ell n\, \ell(r)$ can be assumed to be normal, they can be obtained immediately from the mean and standard deviation of the logarithm of likelihood ratio.

Let us consider the case in which the incoming signal is a rectangular pulse which is M/W seconds long.* The energy of the pulse is half its duration times the amplitude squared of its envelope, for a normalized circuit impedance of one ohm.

---

* The problem of finding the distribution for the sum of M independent random variables, each with a probability density function $f(x) = x \exp\left[-(1/2)(x^2 + a^2)\right] I_o(ax)$ arises in the unpublished report by J. I. Marcum, A Statistical Theory of Target Detection by Pulsed Radar: Mathematical Appendix, Project Rand Report R - 113. Marcum gives an exact expression for this distribution which is useful only for small values of M, and an approximation in Gram-Charlier series which is more accurate than the normal approximation given here. Marcum's expressions could be used in this case, and in the case presented in Section 4.6.

Thus of the WT numbers $\{f_i\}$, there are M consecutive ones which are not zero. These are given by

$$f_i = \sqrt{\frac{2EW}{M}} \quad , \tag{132}$$

where E is the pulse energy at point A in Fig. 5 in the absence of noise. For this case, Eq. (130) and Eq. (131) become

$$\begin{aligned} \mu_{SN}\left(\ell n\ \ell(r)\right) &= \frac{1}{M}\,\frac{E^2}{N_o{}^2} \quad , \\ \mu_{N}\left(\ell n\ \ell(r)\right) &= 0 \quad , \\ \sigma_{SN}{}^2\left(\ell n\ \ell(r)\right) &= \frac{E^2}{MN_o{}^2}\left(1+\frac{2}{M}\frac{E}{N_o}\right) , \end{aligned} \tag{133}$$

and

$$\sigma_{N}{}^2\left(\ell n\ \ell(r)\right) = \frac{E^2}{MN_o{}^2} \quad .$$

The distribution of $\ell n\ \ell(r)$ is approximately normal if M is much larger than one, for, by the central limit theorem, the distribution of a sum of M independent random variables with a common distribution must approach the normal distribution as M becomes large. The actual distribution for the case of noise alone can be calculated in this case, since the convolution integral for the $g_N(r_i)$ with itself any number of times can be expressed in closed form. The distribution of $\ell n\ \ell(r)$ for signal plus noise is more nearly normal than its distribution with noise alone, since the distributions $g_{SN}(r_i)$ are more nearly normal than $g_N(r_i)$.

The receiver operating characteristic for the case M = 16 is plotted in Fig. 6 using the normal distribution as approximation to the true distribution. In many cases it will be found that

$$\frac{1}{M}\cdot\frac{2E}{N_o} << 1 \quad . \tag{134}$$

In such a case the distributions have approximately the same variance. Assuming normal distribution then leads to the curves of Figs. 2 and 3, with

$$d = \frac{1}{4M}\left(\frac{2E}{N_o}\right)^2 \quad . \tag{135}$$

4.8 A Radar Case

This section deals with detecting a radar target at a given range. That is, we shall assume that the signal, if it occurs, consists of a train of M pulses whose time of occurrence and envelope shape are known. The carrier phase will be assumed to have a uniform distribution for each pulse independent of all others, i.e., the pulses are incoherent.

The set of signals can be described as follows:

$$s(t) = \sum_{m=0}^{M-1} f(t+m\tau)\cos\left(\omega t+\theta_i\right) , \tag{136}$$

where the M angles θ_i have independent uniform distributions, and the function f, which is the envelope of a single pulse, has the property that

$$\int_0^T f(t+i\tau)\, f(t+j\tau)\, dt = \frac{2E}{M}\,\delta_{ij} \quad , \tag{137}$$

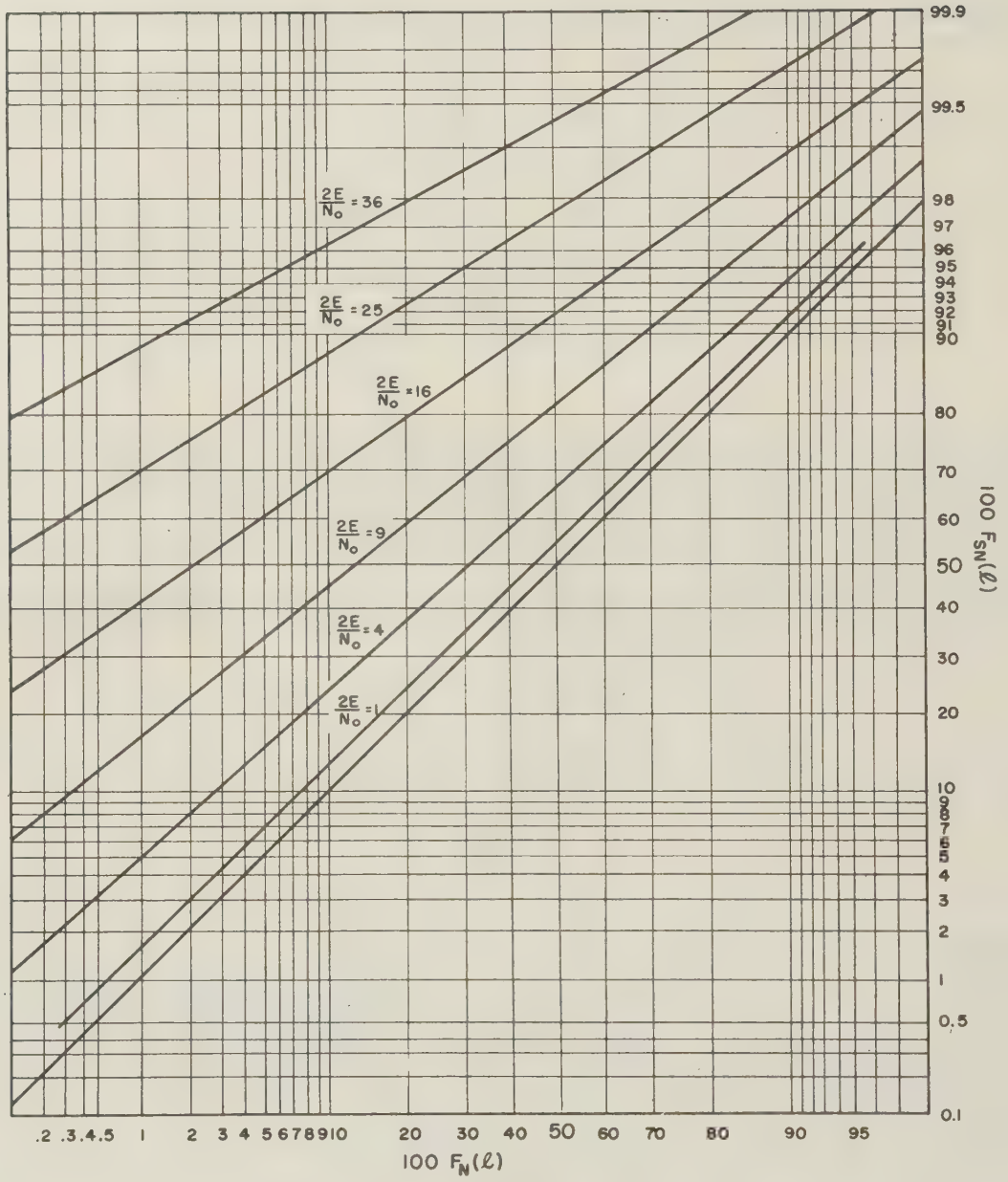


Fig. 6

RECEIVER OPERATING CHARACTERISTIC

BROAD BAND RECEIVER WITH
OPTIMUM VIDEO DESIGN, M = 16

where δ_{ij} is the Kronecker delta function, which is zero if $i \neq j$, and unity if $i = j$. The time τ is the interval between pulses. Eq. (137) states that the pulses are spaced far enough so that they are orthogonal, and that the total signal energy is E.* The function f(t) is also assumed to have no frequency components as high as $\omega/2\pi$.

The likelihood ratio can be obtained by applying Eq. (56). Then

$$\ell(x) = \int_R \exp\left[-\frac{E(s)}{N_o}\right] \exp\left[\frac{2}{N_o}\int_0^T s(t)\, x(t)\, dT\right] dP_S(s) \tag{138}$$

or

$$\ell(x) = \exp\left[-\frac{E}{N_o}\right] \int_0^{2\pi} \cdots \int_0^{2\pi} \exp\left[\frac{2}{N_o}\int_0^T \sum_{m=o}^{M-1} f(t+m\tau)x(t)\cos(\omega t+\theta_m)dt\, d\theta_o \ldots d\theta_{M-1}\right] . \tag{139}$$

The integral can be evaluated, as in Section 4.5, yielding

$$\ell(x) = \exp\left[-\frac{E}{N_o}\right] \prod_{m=0}^{M-1} I_o\left(\frac{r_m}{N}\right) , \tag{140}$$

where

$$\left(\frac{r_m}{N}\right)^2 = \left[\frac{2}{N_o}\int_0^T f(t+m\tau)x(t)\cos\omega t dt\right]^2 + \left[\frac{2}{N_o}\int_0^T f(t+m\tau)\, x(t)\sin\omega t dt\right]^2 . \tag{141}$$

This quantity r_m is almost identical with the quantity r which appeared in the discussion of the case of the signal known except for carrier phase, Section 4.5. In fact, each r_m could be obtained in a receiver in the manner described in that section. The quantity r_o is connected with the first pulse; it could be obtained by designing an ideal filter for the signal

$$s_o(t) = f(t)\cos(\omega t+\theta) \tag{142}$$

for any value of the phase angle θ, and putting the output through a linear detector. The output will be $(N_o/2)r_o/N$ at some instant of time t_o which is determined by the time delay of the filter. The other quantities r_m differ only in that they are associated with the pulses which come later. The output of the filter at time $t_o + m\tau$ will be $(N_o/2)r_m/N$.

It is convenient to have the receiver calculate the logarithm of the likelihood ratio,

$$\ln \ell(x) = -\frac{E}{N_o} + \sum_{m=o}^{M-1} \ln I_o\left(\frac{r_m}{N}\right) . \tag{143}$$

Thus the $\ln I_o (r_m/N)$ must be found for each r_m, and these M quantities must be added. As in the previous section, r_m/N will usually be small enough so that $\ln I_o (x)$ can be approximated by $x^2/4$. The quantities $1/4 (r_m/N)^2$ can be found by using a square law detector rather than a linear detector, and the outputs of the square law detector at times t_o, $t_o + \tau$, ... , $t_o + (M-1)\tau$ then must be added. The ideal system thus consists of an IF amplifier with its passband matched to a single pulse,** a

---

* The factor 2 appears in (137) because f(t) is the pulse envelope; the factor M appears because the total energy E is M times the energy of a single pulse.

** It is usually most convenient to made the ideal filter (or an approximation to it) a part of the IF amplifier

square law detector (for the threshold signal case), and an integrating device.

We shall find normal approximations for the distribution functions of the logarithm of the likelihood ratio using the approximation

$$\ell n\, I_o\left(\frac{r_m}{N}\right) \approx \frac{r_m^2}{4N^2} \tag{144}$$

which is valid for small values of r_m/N.* Substitution of (144) into (143) yields

$$\ell n\, \ell \approx -\frac{E}{N_o} + \sum_{n=o}^{M-1} \frac{1}{4}\left(\frac{r_m}{N}\right)^2 . \tag{145}$$

The distributions for the quantities r_m are independent; this follows from the fact that the individual pulse functions $f(t+m\tau)\cos(\omega t+\theta_m)$ are orthogonal. The distribution for each is the same as the distribution for the quantity r which appears in the discussion of the signal known except for phase; the same analysis applies to both cases. Thus, by Eq. (83)**

$$P_N\left(\frac{r_m}{N}\sqrt{\frac{N_o M}{2E}} \geq \alpha\right) = \exp\left[-\frac{\alpha^2}{2}\right]$$

$$P_N\left(\frac{r_m}{N} \geq a\right) = \exp\left[-\frac{a^2 N_o M}{2E}\right] , \tag{146}$$

and by (89),

$$P_{SN}\left(\sqrt{\frac{N_o M}{2E}}\,\frac{r_m}{N} \geq \alpha\right) = \exp\left[-\frac{E}{N_o}\right]\int_\alpha^\infty \alpha \exp\left[-\frac{\alpha^2}{2}\right] I_o\left(\alpha\sqrt{\frac{2E}{N_o M}}\right) d\alpha \tag{147}$$

or

$$P_{SN}\left(\frac{r_m}{N} \geq a\right) = \frac{N_o M}{2E}\exp\left[-\frac{E}{N_o M}\right]\int_a^\infty a \exp\left(-\frac{a^2 N_o M}{4E}\right) I_o(a)\, da .$$

The density functions can be obtained by differentiating (146) and (147):

$$g_N\left(\frac{r_m}{N}\right) = \frac{MN_o}{2E}\left(\frac{r_m}{N}\right)\exp\left[-\left(\frac{r_m}{N}\right)^2\left(\frac{N_o M}{4E}\right)\right] ,$$

$$g_{SN}\left(\frac{r_m}{N}\right) = \frac{MN_o}{2E}\left(\frac{r_m}{N}\right)\exp\left[-\frac{E}{MN_o}\right]\exp\left[-\left(\frac{r_m}{N}\right)^2\left(\frac{N_o M}{4E}\right)\right] I_o\left(\frac{r_m}{N}\right). \tag{148}$$

---

* See the footnote below equation (131).

** The M appears in the following equations because the energy of a single pulse is E/M rather than E.

This is the same situation, mathematically, as appeared in the previous section. The standard deviation and the mean for the logarithm of the likelihood ratio can be found in the same manner, and they are

$$\mu_{SN}(\ell n\, \ell) = \frac{E^2}{MN_o^2} \quad ,$$

$$\mu_{N}(\ell n\, \ell) = 0 \quad ,$$

$$\sigma_{SN}^2(\ell n\, \ell) = \frac{E^2}{MN_o^2}\left(1 + \frac{2E}{MN_o}\right),$$

and

$$\sigma_{N}^2(\ell n\, \ell) = \frac{E^2}{MN_o^2} \quad . \tag{149}$$

If the distributions can be assumed normal, they are completely determined by their means and variances. These formulas are identical with the formulas (133) of the previous section. The problem is the same, mathematically, and the discussion and receiver operating characteristic curves at the end of Section 4.7 apply to both cases.

4.9 Approximate Evaluation of an Optimum Receiver

In order to obtain approximate results for the remaining two cases, the assumption is made that in these cases the receiver operating characteristic can be approximated by the curves of Figs. 2 and 3, i.e., that the logarithm of the likelihood ratio is approximately normal. This section discusses the approximation and a method for fitting the receiver operating characteristic to the curves of Figs. 2 and 3.

By (68), $F_{SN}(\ell)$ can be calculated if $F_N(\ell)$ is known. Furthermore, it can be seen that the n^{th} moment of the distribution $F_N(\ell)$ is the $(n-1)^{th}$ moment of the distribution $F_{SN}(\ell)$. Hence, the mean of the likelihood ratio with noise alone is unity, and if the variance of the likelihood ratio with noise alone is σ_N^2, the second moment with noise alone, and hence the mean with signal plus noise, is $1 + \sigma_N^2$. Thus the difference between the means is equal to σ_N^2, which is the variance of the likelihood ratio with noise alone. Probably this number characterizes ability to detect signals better than any other single number.

Suppose the logarithm of the likelihood ratio has a normal distribution with noise alone, i.e.,

$$F_N(\ell) = \frac{1}{\sqrt{2\pi d}} \int_{\ell n\, \ell}^{\infty} \exp\left[-\frac{(x-m)^2}{2d}\right] dx, \tag{150}$$

where m is the mean and d the variance of the logarithm of the likelihood ratio. The n^{th} moment of the likelihood ratio can be found as follows:

$$\mu_N(\ell^n) = \int_0^{\infty} \ell^n\, dF_N(\ell) = \frac{1}{\sqrt{2\pi d}} \int_{-\infty}^{\infty} \exp[nx]\exp\left[-\frac{(x-m)^2}{2d}\right] dx \quad , \tag{151}$$

where the substitution $\ell = \exp x$ has been made. The integral can be evaluated by completing the square in the exponent and using the fact that

$$\int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{2d}\right] dx = \sqrt{2\pi d} \quad .$$

Thus

$$\mu_N(\ell^n) = \exp\left[\frac{n^2 d}{2} + nm\right] \quad . \tag{152}$$

In particular, the mean of $\ell(x)$, which must be unity, is

$$\mu_N(\ell) = 1 = \exp\left[\frac{d}{2} + m\right] \quad , \tag{153}$$

and therefore

$$m = -\frac{d}{2} \quad . \tag{154}$$

The variance of $\ell(x)$ with noise alone is σ_N^2, and therefore the second moment of $\ell(x)$ is

$$\mu_N(\ell^2) = \left[\mu_N(\ell)\right]^2 + \sigma_N^2(\ell) = 1 + \sigma_N^2(\ell) \quad , \tag{155}$$

and this must agree with (152). It follows that

$$\mu_N(\ell^2) = 1 + \sigma_N^2 = \exp\left[2d + 2m\right] = \exp\left[d\right] \tag{156}$$

and therefore

$$d = \ell n\ (1 + \sigma_N^2)$$

The distribution of likelihood ratio with signal plus noise can be found by applying Eq. (68). Thus

$$\begin{aligned} dF_{SN}(\ell) &= \ell\, dF_N(\ell) \quad , \\ F_{SN}(\ell) &= -\int_{\ell}^{\infty} \ell\, dF_N(\ell) \end{aligned} \tag{158}$$

If $dF_N(\ell)$ is obtained from Eq. (150) and ℓ is replaced by exp x, then

$$F_{SN}(\ell) = \frac{1}{\sqrt{2\pi d}} \int_{\ell n \ell}^{\infty} \exp\left[x\right] \exp\left[-\frac{(x + \frac{d}{2})^2}{2d}\right] dx$$

or

$$F_{SN}(\ell) = \frac{1}{\sqrt{2\pi d}} \int_{\ell n \ell}^{\infty} \exp\left[-\frac{(x - \frac{d}{2})^2}{2d}\right] dx \quad . \tag{159}$$

Thus the distribution of $\ell n \ell$ is normal also when there is signal plus noise, in this case with mean d/2 and variance d.

In summary, it is probable that the variance σ_N^2 of the likelihood ratio measures ability to detect signals better than any other single number. If the logarithm of likelihood ratio has a normal distribution with noise alone, then this distribution and that with signal plus noise are completely determined if σ_N^2 is given. The distribution of $\ell n\ \ell(x)$ is normal in both cases. Its variance in both cases is d, which is also the difference of the means. The receiver operating characteristic curves are those plotted in Fig. 2, with the parameter d related to σ_N^2 by the equation

$$d = \ell n\ (1 + \sigma_N^2) \tag{160}$$

In the case of a signal known exactly, this is the distribution which occurs. In the cases of Section 4.6, Section 4.7, and Section 4.8 this distribution is found to be the limiting distribution when the number of sample points is large. Certainly in most cases the distribution has this general form. Thus it seems reasonable that useful approximate results could be obtained by calculating only σ_N^2 for a given case and assuming that the ability to detect signals is approximately the same as if the logarithm of the likelihood ratio had a normal distribution. On this basis, $\sigma_N^2\ (\ell)$ is calculated in the following sections for two cases, and the assertion is made that the receiver operating characteristic curves are approximated by those of Fig. 2 with $d = \ell n\ (1 + \sigma_N^2)$.

4.10 Signal Which is One of M Orthogonal Signals

Suppose that the set of expected signals includes just M functions $s_k(t)$, all of which have the same probability, the same energy E, and are orthogonal. That is,

$$\int_0^T s_k(t)\, s_q(t)\, dt = E\,\delta_{kq}. \tag{161}$$

Then the likelihood ratio can be found from Eq. (56) to be

$$\ell(x) = \sum_{k=1}^{M} \frac{1}{M} \exp\left[-\frac{E}{N_o}\right] \exp\left[\frac{1}{N}\sum_{i=1}^{n} x_i s_{ki}\right],$$

or

$$\ell(x) = \frac{1}{M}\sum_{k=1}^{M} \exp\left[\frac{1}{N}\sum_{i=1}^{n} x_i s_{ki} - \frac{E}{N_o}\right], \tag{162}$$

where s_{ki} are the sample values of the function $s_k(t)$.

With noise alone, each term of the form $(1/N)\sum_{i=1}^{n} x_i s_{ki}$ has a normal distribution with mean zero and variance $\sum_{i=1}^{n} s_{ki}^2/N = 2E/N_o$.* Furthermore, the M different quantities $(1/N)\sum_{i=1}^{n} x_i s_{ki}$ are independent, since the functions $s_k(t)$ are orthogonal. It follows that the terms $\exp\left[(1/N)\sum_{i=1}^{n} x_i s_{ki} - E/N_o\right]$ are independent.

Since the logarithm of each term $Z = \exp\left[(1/N)\sum_{i=1}^{n} x_i s_{ki} - E/N_o\right]$ has a normal distribution with mean $(-E/N_o)$ and variance $2E/N_o$, the moments of the distribution can be found from Eq. (152). The n^{th} moment is

$$\mu_N(Z^n) = \exp\left[n(n-1)\frac{E}{N_o}\right]. \tag{163}$$

It follows that the mean of each term is unity, and the variance is

$$\sigma_N^2(Z) = \mu(Z^2) - \left[\mu(Z)^2\right] = \exp\left[\frac{2E}{N_o}\right] - 1. \tag{164}$$

The variance of a sum of independent random variables is the sum of the variances of the terms. Therefore

$$\sigma_N^2(M\ell) = M\left[\exp\left(\frac{2E}{N_o}\right) - 1\right], \tag{165}$$

and it follows that the variance of the likelihood ratio is

$$\sigma_N^2(\ell) = \frac{1}{M}\left[\exp\left(\frac{2E}{N_o}\right) - 1\right]. \tag{166}$$

It was pointed out in Section 4.9, that the receiver operating characteristic curves are approximately those of Fig. 2, with

$$d = \ell n\,(1 + \sigma_N^2) = \ell n\left(1 - \frac{1}{M} + \frac{1}{M}\exp\left(\frac{2E}{N_o}\right)\right). \tag{167}$$

---

* The reasoning is the same as that in Section 4.4.

This equation can be solved for $2E/N_o$:

$$\frac{2E}{N_o} = \ell n \left[1 + M (e^d - 1) \right] \quad . \tag{168}$$

Suppose it is desired to keep the false alarm probability and probability of detection constant. This requires that d be kept constant. Then from Eq. (168) it can be seen that if the number of possible signals M is increased, the signal energy E must also be increased.

4.11 Signal Which is One of M Orthogonal Signals with Unknown Carrier Phase

Consider the case in which the set of expected signals includes just M different amplitude-modulated signals which are known except for carrier phase. Denote the signals by

$$s_k(t) = f_k(t) \cos (\omega t + \theta) \quad . \tag{169}$$

It will be assumed further that the functions $f_k(t)$ all have the same energy E and are orthogonal, i.e.,

$$\int_0^T f_k(t)\, f_q(t)\, dt = 2E \delta_{kq} \quad , \tag{170}$$

where the 2 is introduced because the f's are the signal amplitudes, not the actual signal functions. Also, let the $f_k(t)$ be band-limited to contain no frequencies as high as . Then it follows that any two signal functions with different envelope functions will be orthogonal. Let us assume also that the distribution of phase, θ, is uniform, and that the probability for each envelope function is 1/M.

With these assumptions, the likelihood ratio can be obtained from Eq. (66), and it is given by

$$\ell(x) = \frac{1}{M} \sum_{k=1}^{M} \frac{1}{2\pi} \int_0^{2\pi} \exp\left[\frac{1}{N} \sum_{i=1}^{n} x_i s_{ki} - \frac{E}{N_o} \right] d\theta$$

where s_{ki} are the sample values of $s_k(t)$, and hence depend upon the phase θ. The integration is the same as in the case of the signal known except for phase, and the result, obtained from Eq. (82), is

$$\ell(x) = \frac{1}{M} \sum_{k=1}^{M} \exp\left[- \frac{E}{N_o} \right] I_o\left(\frac{r_k}{N} \right) \quad , \tag{172}$$

where

$$r_k = \sqrt{\left(\sum_i x_i f_k(t_i) \cos \omega t_i \right)^2 + \left(\sum_i x_i f_k(t_i) \sin \omega t_i \right)^2} \quad . \tag{173}$$

Now the problem is to find $\sigma_N^2(\ell)$. The variance of each term in the sum in Eq. (172) can be found since the distribution function with noise alone can be found in Section 4.5. Since the $f_k(t)$ are orthogonal, the distributions of the r_k are independent, and the terms in the sum in Eq. (172) are independent. Then the variance of the likelihood ratio, $\sigma_N^2(\ell)$, is the sum of the variances of the terms, divided by M^2.

The distribution function for each term $\exp(-E/N_o)\ I_o(r_k/N)$ is given in Section 4.5 by Eqs. (84) and (85). If α is defined by the equation

$$\beta = \exp\left[- \frac{E}{N_o} \right] I_o\left(\alpha \sqrt{\frac{2E}{N_o}} \right) \quad , \tag{174}$$

then the distribution function in the presence of noise for each term in Eq. (172) is

$$F_N^{(k)}(\beta) = \exp\left[-\frac{\alpha^2}{2}\right] . \tag{175}$$

The mean value of each term is

$$\mu_N^{(k)}(\beta) = \int_0^\infty \beta dF_N^{(k)}(\beta) = \int_0^\infty \exp\left[-\frac{E}{N_o}\right] I_o\left(\sqrt{\frac{2E}{N_o}}\,\alpha\right)\alpha \exp\left[-\frac{\alpha^2}{2}\right] d\alpha . \tag{176}$$

This can be evaluated as on page 174 of Threshold Signals[5], and the result is that $\mu^{(k)}(\beta) = 1$.
The second moment of each term is

$$\mu_N^{(k)}(\beta^2) = \int_0^\infty \beta^2\, dF_N^{(k)}(\beta) ,$$

or

$$\mu_N^{(k)}(\beta^2) = \int_0^\infty \exp\left[-\frac{2E}{N_o}\right]\left[I_o\left(\alpha\sqrt{\frac{2E}{N_o}}\right)\right]^2 \alpha \exp\left[-\frac{\alpha^2}{2}\right] d\alpha . \tag{177}$$

The integral can be evaluated as in Appendix E of Part II of reference 17, and the result is

$$\mu_N^{(k)}(\beta^2) = I_o\left(\frac{2E}{N_o}\right) . \tag{178}$$

The variance of each term in Eq. (172) is

$$\left[\sigma_N^{(k)}(\beta)\right]^2 = \mu^{(k)}(\beta^2) - \left[\mu^{(k)}(\beta)\right]^2 = I_o\left(\frac{2E}{N_o}\right) - 1 . \tag{179}$$

It follows that the variance of $M\ell$ is

$$\sigma_N^2(M\ell) = M\left[I_o\left(\frac{2E}{N_o}\right) - 1\right] , \text{ and therefore} \tag{180}$$

$$\sigma_N^2(\ell) = \frac{1}{M}\left[I_o\left(\frac{2E}{N_o}\right) - 1\right] , \tag{181}$$

since the variance for the sum of independent random variables is the sum of the variances.

If the approximation described in Section 4.9 is used, the receiver operating characteristic curves are approximately those of Fig. 2, with

$$d = \ell n\,(1 + \sigma_N^2) = \ell n\left(1 - \frac{1}{M} + \frac{1}{M} I_o\left(\frac{2E}{N_o}\right)\right). \tag{182}$$

4.12 The Broad Band Receiver and the Optimum Receiver

A few applications of the results of Section 4 are suggested in Table I, Section 4.1. Two further examples of practical knowledge obtainable from the theory are presented in this section and in the next.

One common method of detecting pulse signals in a frequency band of width B is to build a receiver which covers this entire frequency band. Such a receiver with a pulse signal of known starting time is studied in Section 4.7. This is not a truly optimum receiver; it would be interesting to compare it with an optimum receiver. We have been unable to find the distribution of likelihood ratio for the case of a signal which is a pulse of unknown carrier phase if the frequency is distributed evenly over a band. However, if the problem is changed slightly, so that the frequency is restricted to points spaced approximately the reciprocal of the pulse width apart, then pulses at different frequencies are approximately orthogonal, and the case of the signal which is one of M orthogonal signals known except for phase can be applied. Eq. (182) should be used with M equal to the ratio of the frequency band width B to the pulse band width. Since the band width of a pulse is approximately the reciprocal of its pulse width, the parameter M used in Section 4.7 also has this value. Curves showing $2E/N_o$ as a function of d are given in Fig. 7 for both the approximate optimum receiver and the broad band receiver for several values of M. In the figure, d is calculated from Eq. (135) and Eq. (182), which hold for large values of M.

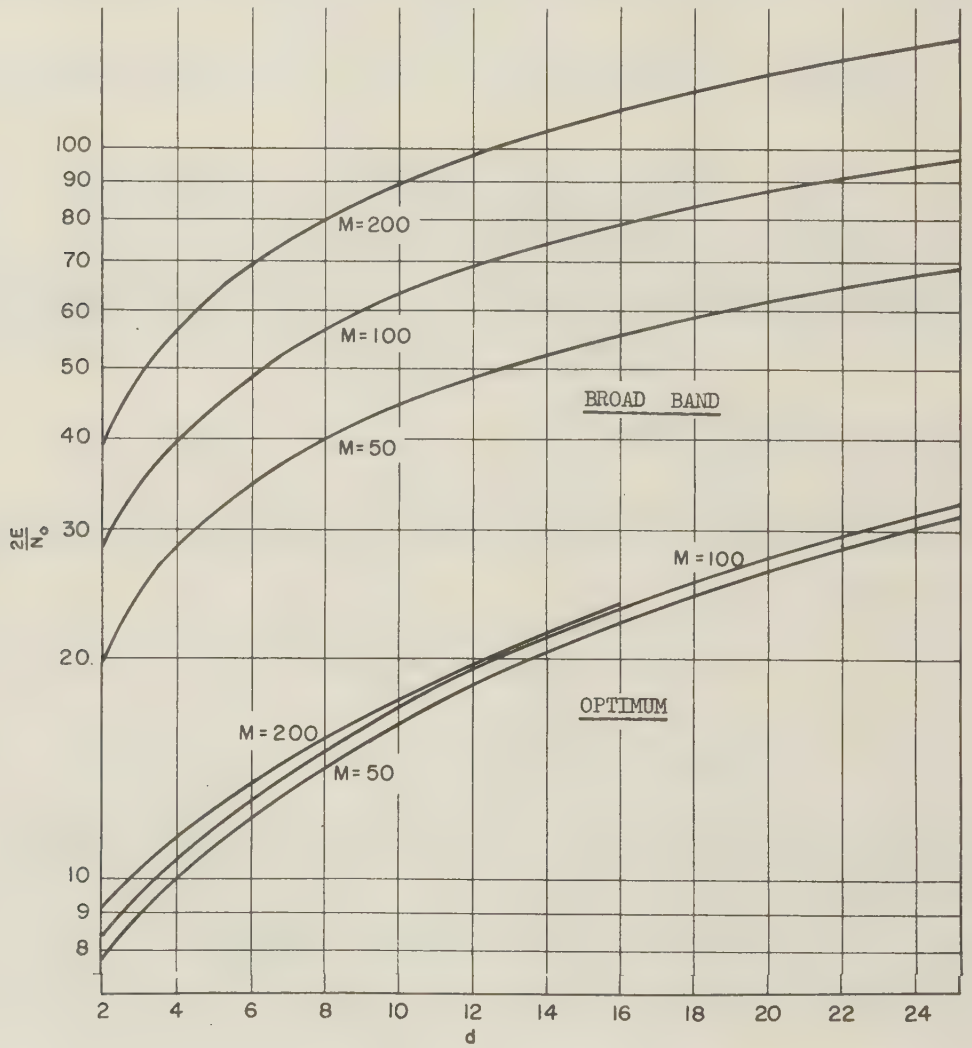


FIG. 7 COMPARISON OF OPTIMUM AND BROAD BAND RECEIVERS

4.13 Uncertainty and Signal Detectability

In the two cases where the signal considered is one of M orthogonal signals, the uncertainty of the signal is a function of M. This provides an opportunity to study the effect of uncertainty on signal detectability. In the approximate evaluation of the optimum receiver when the signal is one of M orthogonal functions, the ROC curves of Figs. 2 and 3 are used with the detection index d given by

$$d = \ell n \left[1 - \frac{1}{M} + \frac{1}{M} \exp\left(\frac{2E}{N_o}\right) \right]. \tag{167}$$

This equation can be solved for the signal energy, yielding

$$\frac{2E}{N_o} = \ell n \left[1 - M + Me^d \right] \approx \ell n \; M + \ell n \; (e^d - 1) \quad , \tag{175}$$

the approximation holding for large $2E/N_o$.* From this equation it can be seen that the signal energy is approximately a linear function of ℓn M when the detection index d, and hence the ability to detect signals, is kept constant. It might be suspected that $2E/N_o$ is a linear function of the entropy, $-\sum p_i \ell n p_i$, where p_i is the probability of the i^{th} orthogonal signal. The linear relation holds only when all the p_i are equal. The expression which occurs in this more general case is:

$$\frac{2E}{N_o} \approx - \ell n \left[\sum p_i^2 \right] + \ell n \; (e^d - 1) \, . \tag{176}$$

LIST OF REFERENCES

1. S. Goldman, Information Theory, Prentice-Hall, New York, 1953. Chapter II, pp. 65-84, is devoted to sampling plans.

2. C. E. Shannon, "Communication in the Presence of Noise," Proc. I.R.E., Vol. 37, pp. 10-21, January, 1949.

3. U. Grenander, "Stochastic Processes and Statistical Inference," Arkiv För Mathematik, Bd 1 nr 17, p. 195, 1950.

4. J. Neyman, and E. S. Pearson, "On the Problems of the Most Efficient Tests of Statistical Hypotheses," Philosophical Transactions of the Royal Society of London, Vol. 231, Series A, p. 289, 1933.

5. J. L. Lawson, and G. E. Uhlenbeck, Threshold Signals, McGraw-Hill, New York, 1950.

6. P. M. Woodward and I. L. Davies, "Information Theory and Inverse Probability in Telecommunications," Proc. I.E.E. (London), Vol. 99, Part III, pp. 37-44, March, 1952.

7. I. L. Davies, "On Determining the Presence of Signals in Noise," Proc. I.E.E. (London), Vol. 99, Part III, pp. 45-51, March, 1952.

8. A. Wald, Sequential Analysis, John Wiley and Sons, 1947.

9. W. C. Fox, "Signal Detectability: A Unified Description of Statistical Methods Employing Fixed and Sequential Observation Processes," Electronic Defense Group, University of Michigan, Technical Report No. 19 (unclassified).

---

* If $2E/N_o > 3$, the error is less than 10%.

10. A. Wald and J. Wolfowitz, "Optimum Character of the Sequential Probability Ratio Test," Ann. Math. Stat., Vol. 19, p. 326, September, 1948.

11. E. Reich, and P. Swerling, "The Detection of a Sine Wave in Gaussian Noise," Journal Applied Physics, Vol. 24, p. 289, March, 1953.

12. R. C. Davis, "On the Detection of Sure Signals in Noise," Journal Applied Physics, Vol. 25, pp. 76-82, January, 1954.

13. J. V. Harrington, and T. F. Rogers, "Signal-to-Noise Improvement Through Integration in a Storage Tube," Proc. I.R.E., Vol. 38, p. 1197, October, 1950.

A. E. Harting, and J. E. Meade, "A Device for Computing Correlation Functions," Rev. Sci. Instr., Vol. 23, 347, 1952.

Y. W. Lee, T. P. Cheatham, Jr., and J. B. Wiesner, "Applications of Correlation Analysis to the Detection of Periodic Signals in Noise," Proc. I.R.E., Vol. 38, p. 1165, October, 1950.

M. J. Levin, and J. F. Reintjes, "A Five Channel Electronic Analog Correlator," Proc. Nat. El. Conf., Vol. 8, 1952.

14. D. O. North, "An Analysis of the Factors which Determine Signal-Noise Discrimination in Pulsed Carrier Systems," RCA Laboratory Rpt PTR-6C, 1943.
See also Reference 5, p. 206.

15. Graphs of values of the integral (89) along with approximate expressions for small and for large values of appear in Rice, S. O., "Mathematical Analysis of Random Noise," B.S.T.J., Vol. 23, p. 282-332 and Vol. 24, p. 46-156, 1944-5. Tables of this function have been compiled by J. I. Marcum in an unpublished report of the Rand Corporation, "Table of Q-Functions," Project Rand Report RM-399.

16. P. G. Hoel, Introduction to Mathematical Statistics, New York: Wiley, 1947, p. 246.

17. The material of Sections 2 and 3 of this paper is drawn from reference 9 above and from Part I of W. W. Peterson, and T. G. Birdsall, "The Theory of Signal Detectability," Electronic Defense Group, University of Michigan, Technical Report No. 13 (Unclassified), July, 1953. Part II of that report contains the material in Section 4 of this paper. Other work in this field may be found in D. Middleton, "Statistical Criteria for the Detection of Pulsed Carriers in Noise," Jour. App. Phys., Vol. 24, p. 371, April, 1953; D. Middleton, "The Statistical Theory of Detection. I: Optimum Detection of Signals in Noise," M.I.T. Lincoln Laboratory, Technical Report No. 35, November 2, 1953; D. Middleton, "Statistical Theory of Signal Detection," Trans. I.R.E., PGIT-3, p. 26, March, 1954; D. Middleton, W. W. Peterson, and T. G. Birdsall, "Discussion of 'Statistical Criteria for the Detection of Pulsed Carriers in Noise. I, II'", Journal Applied Physics, Vol. 25, pp. 128-130, January, 1954.

THE HUMAN USE OF INFORMATION

I. SIGNAL DETECTION FOR THE CASE OF THE SIGNAL KNOWN EXACTLY

Wilson P. Tanner Jr. and John A. Swets

University of Michigan

Abstract

A theory of visual detection is developed, based on the model provided by the theory of signal detectability,[2] and, more generally, by the theory of statistical decision. Two experiments are reported which test some predictions of the theory for the case of the signal-known-exactly. These experiments demonstrate that the human observer tends toward optimum behavior, where optimum behavior is defined as that behavior which maximizes the expected gain from the decision. Their results show the proportion of correct detections to be dependent upon the proportion of false alarms; they indicate that neural activity is a power function of signal intensity. The data also demand a re-evaluation of the threshhold concept. Predictions are made for the data obtained using two different methods of response, forced-choice and yes-no, and the internal consistency of the theory is demonstrated. The predictions of the theory are compared with contrasting predictions of conventional sensory theory; the data are also related to conventional theory.

Introduction

There is some indication that the theory of statistical decision, or in particular, the theory of statistical inference, constitutes a model of relevance to several aspects of human behavior. When a set of rather reasonable assumptions about neurophysiology is coupled with the assumption that the organism tends toward optimum behavior, the theory of statistical decision permits the specification of behavior in a variety of situations that submit to experimental manipulation. In this paper, experiments are reported which were designed to test the predictions that follow from this model for the behavior of the human observer in a visual detection situation. Since several of these predictions are in conflict with predictions of conventional sensory theory, the conventional theory is reviewed in the next section.

Conventional Sensory Theory

For the present purposes, the most significant aspect of conventional sensory theory is its implication that so-called sensory phenomena are peripherally determined. This point of view is directly related to the concept of a threshold, the notion that if some fixed amount of neural activity in the sensory system is exceeded, a signal is detected by the observer with a probability of unity. In this framework, the observer's decision concerning the existence of a signal is assumed to depend entirely upon whether the threshold is exceeded, and the threshold is assumed to be independent of control by essentially non-sensory variables which might influence the attitude or set of the observer. It is also assumed that the threshold is high enough to be exceeded very rarely by sensory system activity unrelated to the presence of a physical signal. If this view is not explicit, it is, at least, implied by the usual treatment of the data.

The primary data from visual detection experiments are frequencies of detection as a function of the intensity of the light signal. In Fig. 1, the dotted lines represent the form of the results of a hypothetical experiment. Consider first a single dotted line. Any point on the line might represent an experimentally determined point. Conventionally, this point is corrected for chance successes by application of the formula,

$$p = \frac{p' - c}{1 - c} \qquad (1)$$

* This paper is based on work done for the U. S. Army Signal Corps under Contract No. Da - 36 - 039 sc - 15358. The experiments reported herein have been reported previously in a technical report[4] of the Electronic Defense Group of the University of Michigan. These experiments were conducted in the Vision Research Laboratory of the University of Michigan.

where p' is the observed proportion of positive responses, p is the corrected proportion of positive responses, and c is the intercept of the dotted line at zero signal intensity.

The justification for the use of this chance-correction formula depends upon the validity of the assumption of an independence of the events comprising the p and c terms, or the assumption that a "false alarm" is a guess, independent of neural activity in the sensory system relevant to the decision concerning signal existence. This assumption implies, and is implied by, the assumption of a threshold. In this context the solid curve of Fig. 1, the curve onto which each of the dotted curves can be mapped by application of the chance-correction formula, is regarded as a "true" curve; that is, its parameters are assumed to be characteristic of the observer's sensory system.

Statistical Decision as a Model for a Theory of Visual Detection

The Basis for Considering this Model

The relevance of the theory of statistical decision as a model for visual detection is suggested by the very likely assumption that spontaneous neural activity occurs in the human's sensory system. Although direct observation of this activity has been made entirely on infra-human organisms, the data strongly suggest that extrapolation to humans is reasonable. Now, if the problem of detection is the detection of signals (which presumably have randomly distributed neural effects) in the presence of random interference or noise, then the task of the observer is that of testing statistical hypotheses, and the model provided by the theory of statistical decision should aid in describing his behavior. This point of view suggests replacing the concept of a threshold by a concept of a criterion range of acceptance, the extent of which is controlled by the observer in the interests of optimum behavior. In addition, it suggests considering the probability that noise (spontaneous neural activity) alone may reach levels which will be in the criterion of acceptance. Also, in contrast with conventional theory, a dependence is implied between the conditional probability that neural activity when a signal exists is in the criterion, and the conditional probability that neural activity when no signal is present is in the criterion.

Elaboration of the Theory

In this and subsequent sections, a new theory of visual detection is developed, based on the model constituted by the theory of statistical decision. A chronologically intermediate step between the theory of statistical decision and the sensory theory presented here is the theory of signal detectability, developed for theoretical observers, by Peterson, Birdsall, and Fox.[2] The mathematical developments and symbols used below are those of Peterson, Birdsall, and Fox, unless otherwise stated.

The Form and Treatment of Sensory Information. It is supposed that the information relevant to detection is a display of neural activity at the cortical level. In the case under consideration in which a signal is presented at a specified time in a specified spatial location, it is assumed that the observer will place the same restrictions on the relevant display. Thus, if the observer is asked to state whether a signal exists in location A at time B, he is assumed to consider only that information in the neural display which refers to location A at time B.

A judgement concerning the existence of a signal is presumably based upon some measure, x, of neural activity. It is assumed that there exists a statistical relationship between the measure and signal intensity. That is, the more intense the signal, the greater is the average of the measures resulting. Thus, for any signal there is a universe distribution which is, in fact, a sampling distribution. It includes all measures which might result if the signal were repeated and measured an infinite number of times. The mean of this universe distribution is associated with the intensity level of the signal. The variance may be associated with other parameters of the signal such as duration or size, but this is beyond the scope of this paper.

Fig. 2 shows two distributions: N representing the case where noise alone is sampled, that is, no signal exists, and S+N, the case where signal plus noise exists. The N and S+N distributions are assumed to be probability density functions; thus the ordinate is probability density. The mean of N depends upon the constant, prevailing background intensity; the mean of S+N depends on background-plus-signal intensity. The variance of N depends on signal parameters, not background parameters, in the case considered here; that is, where the observer knows a priori that if a signal exists, then it will be a particular signal. From the way the diagram is conceptualized, the greater the measure x, the more likely it is that this sample represents a signal. But one can never be certain. Thus, if an observer is asked if a signal exists, he is assumed to base his judgment on the quantity of neural activity. He makes an observation, and then attempts to decide whether this observation is more representative of N or S+N. His task is, then, the task of testing a statistical hypothesis.

For mathematical convenience, it is assumed that the distributions shown in Fig. 2 are Gaussian, with variances equal for N and all values of S+N. Experimental results suggest that equal variance

is not a true assumption, but the deviations are not so great that the inconvenience of a more precise assumption is justified for the purpose of this analysis. It is also assumed that there is a cut-off point such that any measure of neural activity which exceeds that cut-off is in the criterion; that is, any value exceeding the cut-off is accepted as representing the existence of a signal, and any value less than the cut-off is regarded as representing noise alone. Again, for mathematical convenience, the cut-off point is assumed to be well-defined and stable.

Now, consider the way in which the placing of the cut-off affects behavior in the case of a given signal. In the lower right-hand corner of Figure 3, the distributions N and S+N are reproduced for a value of d' = 1. d' is the difference between the means of N and S+N in terms of the standard deviation of N. The criterion scale is also calibrated in terms of the standard deviation of N. On the abscissa there is $P_N(A)$, the probability that if no signal exists the measure will be in the criterion, and on the ordinate $P_{SN}(A)$, the probability that if a signal exists the measure will be in the criterion. A, in this terminology, symbolizes "acceptance of the hypothesis that a signal exists."

If the cut-off is at $-\infty$, all measures are in the criterion: $P_N(A) = P_{SN}(A) = 1$. At minus one standard deviation $P_N(A) = .84$, $P_{SN}(A) = .98$. At zero, $P_N(A) = .5$, $P_{SN}(A) = .84$. At plus one $P_N(A) = .16$ and $P_{SN}(A) = .5$, and for plus ∞, $P_N(A) = P_{SN}(A) = 0$. Thus, for d' = 1, this is the curve showing possible detections for each false-alarm rate.

The Optimization Assumption. At this point, it is necessary to make an assumption which will permit specification of the behavior expected of the observer. In conventional theory, the assumption of a fixed threshold has made it possible to derive testable predictions. In the theory presented here, where it is assumed that the position of the cut-off point between acceptance and rejection of the existence of a signal is controlled by the observer, it is necessary to define the method of control exerted by the observer on the cut-off point in order to make predictions. As stated above, it is assumed that the observer's behavior tends toward optimum behavior. More specifically, it is assumed that the observer sets the cut-off point at a position that maximizes the expected gain. That is to say, the level of the cut-off is determined so as to maximize an expected payoff in terms of the values of hits and correct rejections, and the costs of false alarms and misses.

Peterson, Birdsall, and Fox[2] have shown that the optimum behavior (in this case, that behavior resulting in maximizing the expected gain) in any given experimental condition may be represented by a point on the curve of Figure 3 where its slope is w, where

$$w = \frac{1 - P(SN)}{P(SN)} \quad \frac{(V_{N \cdot CA} + K_{N \cdot A})}{(V_{SN \cdot A} + K_{SN \cdot CA})} \tag{2}$$

where P(SN) is the a priori probability of signal occurence, $V_{N \cdot CA}$ is the value of a correct rejection, $K_{N \cdot A}$ is the cost of a false alarm, $V_{SN \cdot A}$ is the value of a correct detection, and $K_{SN \cdot CA}$ is the cost of a miss.

Equation (2) can be derived from the expression for the expected value of a decision,

$$EV = V_{SN \cdot A} P(SN \cdot A) + V_{N \cdot CA} P(N \cdot CA) - K_{SN \cdot CA} P(SN \cdot CA) - K_{N \cdot A} P(N \cdot A), \tag{3}$$

by substituting conditional probabilities for the probabilities of joint occurence; e.g., $P(SN)\ P_{SN}(A)$ for $P(SN \cdot A)$. Then, maximizing EV is equivalent to requiring that $P_{SN}(A) - w\ P_N(A)$ be a maximum. The value of w thus defines the optimum criterion. More precisely, the optimum criterion consists of all measures of neural activity with likelihood greater than w; i.e., w is the critical value of the likelihood ratio where likelihood ratio for a particular measure is defined as $f_{SN}(x)/ f_N(x)$, the ratio of the probability density for that measure if there is signal plus noise to the probability density if there is noise alone. It can be seen from Eq. 2, that as P(SN) or $V_{SN \cdot A}$ increases or $K_{N \cdot A}$ decreases, w becomes smaller and it is worthwhile to accept a higher false alarm rate in the interest of achieving a greater percentage of correct decisions.

The Predicted Form of the Data. Figure 4 shows a family of curves of $P_{SN}(A)$ vs. $P_N(A)$ with d' as the parameter. This is to be compared with the predictions of conventional theory shown in Figure 5, with $P_N(A)$ assumed to represent guesses, or spurious responses unrelated to relevant neural activity. For each value of signal intensity, it is assumed that there is a true value of $P_{SN}(A)$ either for $P_N(A) = 0$ or for some very small value. The chance correction should transform each of these to horizontal lines.

Another way of comparing the predictions of this theory with those of conventional theory is to construct curves showing the predicted shape of the psychophysical function. These curves are shown in Fig. 6, where P(A), the probability of acceptance, is plotted as a function of d', for comparison with the curves of Fig. 1. These curves will not correct into the same curve by the application of the chance correction. The shift is horizontal rather than vertical. The dotted portions of the curve

show that we are dealing with only a part of the curve, and thus, in terms of this theory, it is improper to apply a normalizing procedure such as the chance-correction formula to that part of the curve.

The Forced-Choice Method of Response

The preceeding discussion specifies the behavior expected of the observer, in terms both of the theory presented here and conventional theory, when the so-called yes-no method of response is employed. A second method of response has been used in psychophysical experimentation; this method is known as the forced-choice method. In this method, the observer does not report directly on the existence of a signal but is required to indicate detection by correctly identifying some attribute of the signal. In the specific version of the forced-choice method most commonly used, the observer knows that on each trial the signal will occur in one of four short, adjoining time intervals, and he is forced to choose in which of these intervals he believes the signal occurred.

The Predicted Form of Forced-Choice Data. While conventional theory predicts the same form for data collected under yes-no and forced-choice methods, the theory presented here leads to different predictions for the form of the data collected using the two procedures. The predictions stemming from this theory for forced-choice data are, as in the case of yes-no data, based on the assumptions that the observer works with a continuous variable, the measure of neural activity or likelihood ratio, and behaves optimally in terms of available information. Optimal behavior requires that the observer select the interval with the greatest associated value of likelihood ratio. Then the probability that a correct answer P(C) will result for a given value of d', for the four-choice or four-interval situation, is the probability that the one sample from the S+N distribution is greater than the greatest of three samples from the distribution of N. For the four-choice situation,

$$P(C) = \int_{-\infty}^{+\infty} \left[F(x)\right]^3 g(x)\, dx, \qquad (4)$$

where F(x) is the area of N and g(x) is the ordinate of S+N. In Fig. 7, P(C), as determined by the integration, is plotted as a function of d', under the assumption of equal variance of the N and S+N distributions.

Criterion of Internal Consistency

Since the theory predicts a different form of data for the two response procedures, forced-choice and yes-no, and since the predictions for the two situations are based on the same neurological parameters, the existence of an internal consistency check on the theory is implied. The information on which the observer bases his decision is contained in the same neural display in the forced-choice situation as in the yes-no situation, and presumably, the values of d' obtained from the two procedures for any given signal intensity must be the same. Thus, if the values of d' are estimated from the data obtained when one of these methods is employed, these estimates should furnish a basis for predicting the data obtained using the other method if the theory is internally consistent. Or, equivalently, the criterion of internal consistency is satisfied if both sets of data yield the same estimates of d'.

The First Experiment

Procedure

An experiment was conducted to test the internal consistency of the proposed theory, using three University of Michigan sophomores as observers. A series of eight experimental sessions involving the forced-choice procedure was followed by a series of sixteen sessions in which the yes-no method of response was used. All of the experimental sessions employed a circular signal, thirty minutes of visual angle in diameter, with a duration of .01 second, on a ten foot-lambert background. Five intensity values of signal were used in the forced-choice sessions. The four greatest of these, reduced by a .1 fixed filter, were used in the yes-no sessions. Details of the experimental procedure and the laboratory have been published by Blackwell, Pritchard, and Ohmart.[1]

In the first four yes-no sessions, two values of *a priori* probability, P(SN) equal to .8 and .4 were used. The observers were informed of the value of P(SN) before each session. No values or costs were incorporated in these four sessions; they were excluded from the analysis as practice sessions. In the next twelve yes-no sessions, all of the information necessary for the calculation of a w (the best possible decision level) was furnished to the observers (i.e., P(SN) and the various values and costs). While they did not know the formal calculation of w, that they knew the direction of change in the cut-off point indicated by a change in any of the factors involved in the w - equation was indicated by the fact that the obtained values of $P_N(A)$ varied appropriately with changes in the information given them. The values and costs were made real to the observers, for they were actually paid in cash.

Each session the observers realized a bonus of between one and two dollars.

The first four of these sessions each carried the same value of w since the same payoff was maintained and P(SN) was held at .8. A high value of $P_N(A)$, or false-alarm rate, resulted. In the next four sessions with P(SN) held at .8, $K_{N \cdot A}$ and $V_{N \cdot CA}$ were gradually increased from session to session (not within sessions) until $P_N(A)$ dropped to a low value. Then P(SN) was dropped to .4, $K_{N \cdot A}$ and $V_{N \cdot CA}$ were reduced so that for the next session $P_N(A)$ stayed low. The last three sessions involved successive increases in $V_{SN \cdot A}$ and $K_{SN \cdot CA}$, again forcing $P_N(A)$ toward a higher value.

Results

The Internal Consistency Check. The yes-no data obtained from each observer for each value of signal intensity were plotted in the form of scatter diagrams of $P_{SN}(A)$ vs. $P_N(A)$. Comparison of these scatter diagrams with the theoretical curves of Fig. 4 provides an estimate of d' from yes-no data. Each d' estimated in this way is based on 560 observations. Estimates of d' from forced-choice data are made by entering the forced-choice curve (Fig. 7) using the observed proportion of correct responses as an estimate of P(C). The last two forced-choice sessions were used in this analysis; each value of d' estimated from forced-choice data is based on 100 observations.

Figs. 8, 9, and 10 show log d' as a function of log signal intensity for the three observers. In general, the agreement is good. The deviation of the forced-choice points at the top and bottom of the graphs can be explained on the basis of sampling variation. For the third observer, the lowest forced-choice point is off the graph to the right of the line.

The Relationship Between Neural Activity and Signal Intensity. Figs. 8, 9, and 10 point up another difference between conventional sensory theory and the theory presented here. In conventional theory, the assumption is made that the relationship between neural activity and signal intensity is linear. The results obtained from this experiment suggest that neural activity is a power function of signal intensity, a result that is consistent with a more direct type of neuro-physiological data, in particular, the results of electrical recordings from optic nerve fibers.

An External Consistency Check. The results reported above support internal consistency. The theory also turns out to be consistent with the data in the literature, for, when the d' vs. signal intensity function for any one of the observers is used to predict probability of detection as a function of signal intensity in terms of this theory, the result closely approximates a type of curve, a normal ogive, that is frequently reported. Chi-square analyses suggest that approximately fifteen times the ordinary amount of data would be required to distinguish the predicted curve from a normal ogive.

Additional Analyses Suggested by the Theory. According to conventional theory, application of the chance correction should yield corrected values of $P_{SN}(A)$ which are independent of $P_N(A)$, or should yield corrected thresholds in the conventional sense which are independent of $P_N(A)$. Rank-order correlations for the three observers between $P_N(A)$ and corrected thresholds (.30, .71, .67) are highly significant; the combined $p = .0002$. Similar correlations were obtained (.32, .62, .76) between $P_N(A)$ and corrected $P_{SN}(A)$. These results are consistent with the theory presented here.

Another method of comparison is to fit the scatter diagrams of $P_{SN}(A)$ vs. $P_N(A)$ by straight lines. According to conventional theory, these straight lines should intercept the point (1.00, 1.00). Sampling error would be expected to send some of the lines to either side of this point. The four scatter diagrams obtained from each of the three observers are reproduced in Figs. 11, 12, and 13. All twelve of these lines intersect the line $P_{SN}(A) = 1.00$ at values of $P_N(A)$ between 0 and 1.00, approximately in an order which would be predicted if these lines were arcs of the curves $P_{SN}(A)$ vs. $P_N(A)$ as defined by the theory proposed here.

The Second Experiment

A second experiment was conducted to test the theory proposed here and to provide additional basis for selecting between this theory and conventional theory. This experiment was suggested by R. Z. Norman; its results were reported by Swets.[3]

The Rationale for This Experiment

As pointed out above, conventional theory is consistent with the view that the mechanism of detection is one that triggers when the amount of neural activity exceeds a criterion amount, and loses all discrimination among quantities of neural activity that fall short of this amount. Thus, for a (four-choice) forced-choice situation where the observer is required to indicate a second choice as well as a first choice, conventional theory leads to the prediction that, when the first choice is incorrect, the second choice will be correct with a probability of .33, since the second choice is made

from among three intervals presumably on a chance basis. On the other hand, the theory proposed here supposes that the observer works with a variable x (likelihood ratio) that is continuous throughout the range of x, not merely continuous above a critical point. If this is the case, the observer should be able to rank the four values of x associated with the four intervals; then the probability of a correct second choice, given an incorrect first choice, is greater than .33. The relationship between this predicted probability and d' is given by the expression

$$\frac{3\int_{-\infty}^{+\infty}\left[F(x)\right]^2\left[1 - F(x)\right] g(x)\, dx}{1 - \int_{-\infty}^{+\infty}\left[F(x)\right]^3 g(x)\, dx} \qquad (5)$$

where the symbols have the same meaning as in Eq. (4).

Results

Data were collected from four observers, each of whom served in three sessions. Each session included 150 observations for which both a first and second choice were required. The resulting twelve proportions of correct second choices are plotted against d' in Figure 14. Although a single value of signal intensity was used, the values of d' differed sufficiently from one observer to another to provide an indication of the congruence of the data and the predicted functions. (The function predicted for the three-choice (or three-interval) situation is included in Fig. 14 to emphasize that this function is not the same as the predicted function of the probability of a correct second choice, given an incorrect first choice, for the four-choice situation).

A systematic deviation from the prediction of conventional theory clearly exists. Considering the combined data, the proportion of correct second choices is .46. The deviation of this proportion from .33 is highly significant; the χ^2 obtained (43.66) is more than twice the χ^2 (19.0) associated with a probability of .00001. Allowing for the possibility that being required to make a second choice might depress first-choice performance, blocks of 50 observations for which only a first choice was required were alternated with blocks of 50 observations for which both a first and second choice were required. Pooling the data, the proportions of correct first choices for the two conditions are .650 and .651; this difference is obviously not significant.

The systematic deviation of the second-choice data from the function predicted by the theory proposed here may be a result of the inadequacy of the assumption of equal variance for N and all values of S+N. Any assumption involving a constant ratio of mean to standard deviation would result in lower predicted values for proportions of correct second choices. Determining the proportionality of mean and standard deviation leading to the most adequate predictions, however, does not fall within the scope of this paper. It is clear, nonetheless, that the second-choice data tend to confirm the theory proposed in this paper and to differentiate between this theory and conventional theory.

Conclusions

The following conclusions are advanced:

1) The conventional concept of a threshold, or a threshold region, needs re-evaluating in the light of these data.

2) The assumptions underlying the use of the correction for chance successes are rejected on the basis of statistical tests.

3) Change in neural activity is a power function of change in light intensity.

4) The model provided by the theory of statistical decision, and, in more detail, by the mathematical theory of signal detectability, is applicable to the problems of visual detection.

5) The criterion of seeing depends on psychological as well as physiological factors. In these experiments the observers tended to use optimum criteria.

6) The experimental data support the logical connection between forced-choice and yes-no techniques developed by the theory presented here.

7) A measurable false-alarm rate can be, and should be, produced in yes-no psychophysical experiments.

8) The forced-choice procedure, which does not necessitate the determination of a criterion, should be used whenever possible.

List of References

1. Blackwell, H. R., Pritchard, B. S., and Ohmart, J. G. Automatic apparatus for stimulus presentation and recording in visual threshold experiments. J. Opt. Soc. Amer., 44, 1954.
2. Peterson, W. W., and Birdsall, T. G., and Fox, W. C. The theory of signal detectability. Transactions of the I.R.E. Professional Group on Information Theory (this issue).
3. Swets, J. A. An experimental comparison of two theories of visual detection. Unpublished doctoral dissertation, University of Michigan, 1954.
4. Tanner, W. P., Jr., and Swets, J. A. A new theory of visual detection. Technical Report No. 18, Electronic Defense Group, University of Michigan, 1953.

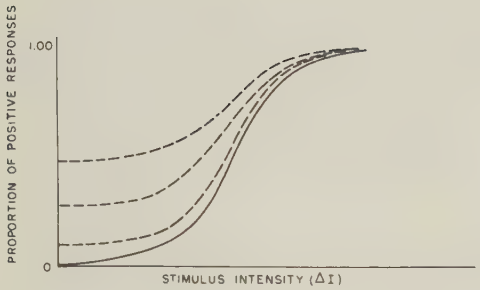


Fig. 1 - Hypothetical data from detection experiments.

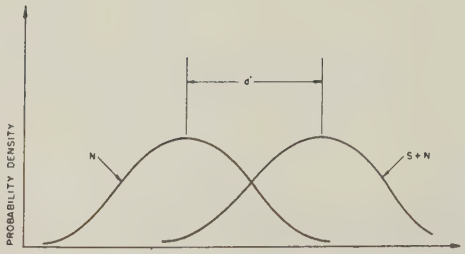


Fig. 2 - Hypothetical distributions of noise and signal plus noise.

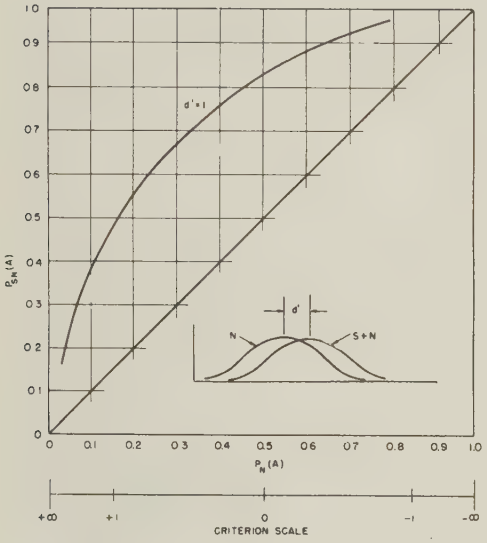


Fig. 3 - $P_{SN}(A)$ vs. $P_N(A)$ for $d' = 1$.

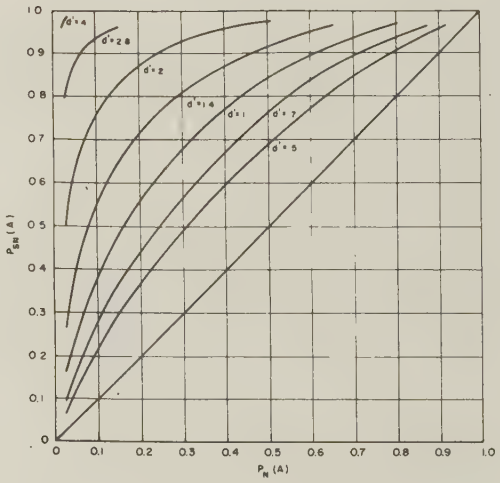


Fig. 4 - $P_{SN}(A)$ vs. $P_N(A)$ with d' as the parameter.

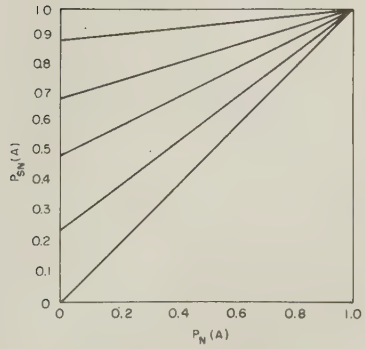


Fig. 5 - $P_{SN}(A)$ vs. $P_N(A)$ as a function of d' assuming conventional theory.

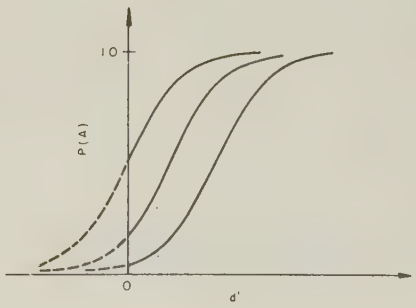


Fig. 6 - P(A) as a function of d' assuming the statistical - decision model.

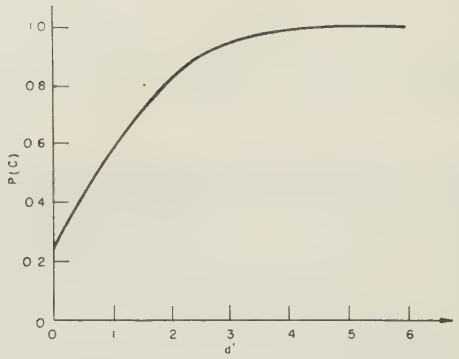


Fig. 7 - P(C) as a function of d' a theoretical curve.

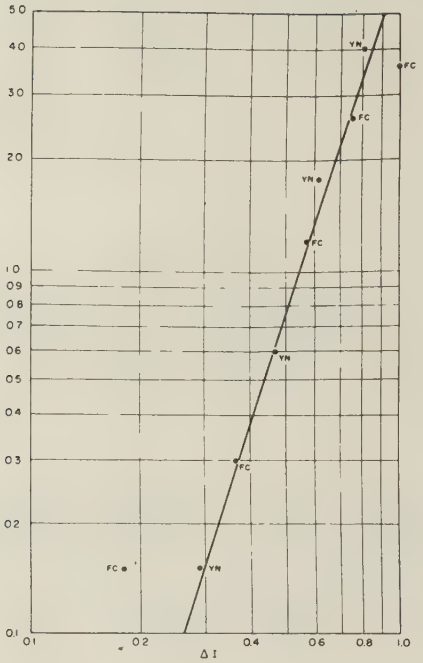


Fig. 8 - Log d' vs. log signal intensity for observer 1.

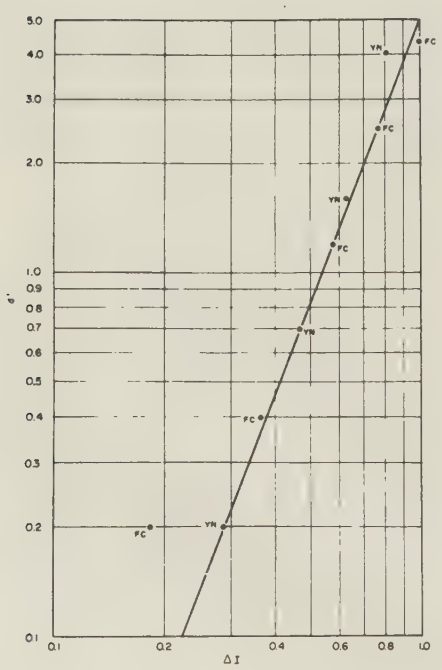


Fig. 9 - Log d' vs. log signal intensity for observer 2.

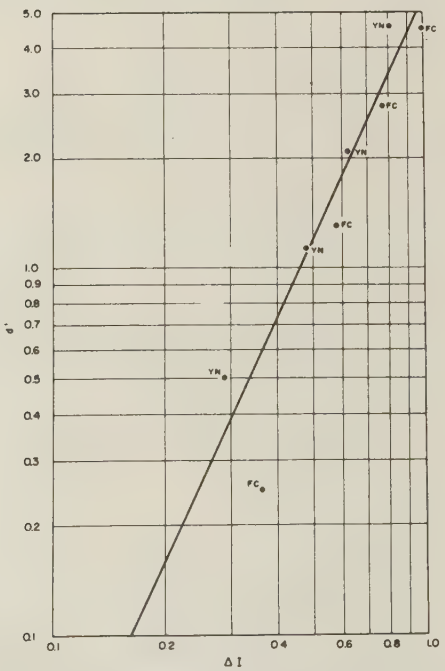


Fig. 10 - Log d' vs. log signal intensity for observer 3.

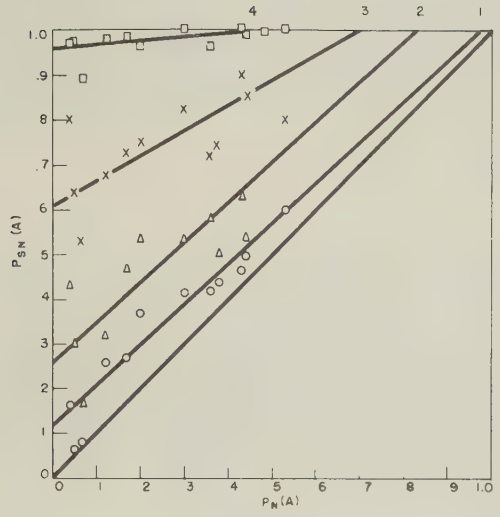


Fig. 11 - $P_{SN}(A)$ vs. $P_N(A)$ for observer 1.

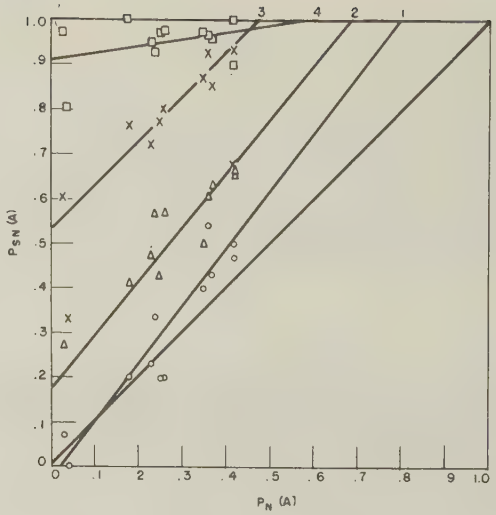


Fig. 12 - $P_{SN}(A)$ vs. $P_N(A)$ for observer 2.

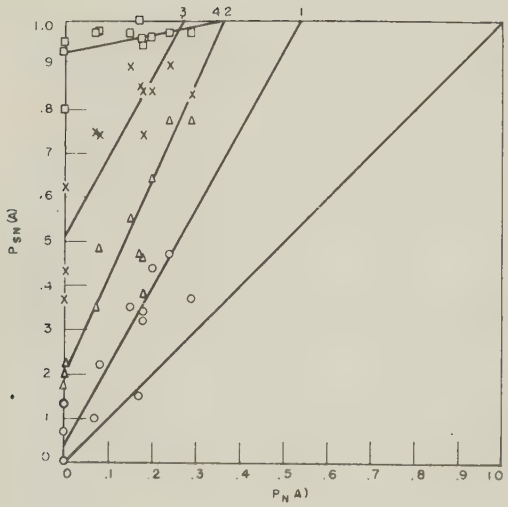


Fig. 13 - $P_{SN}(A)$ vs. $P_N(A)$ for observer 3.

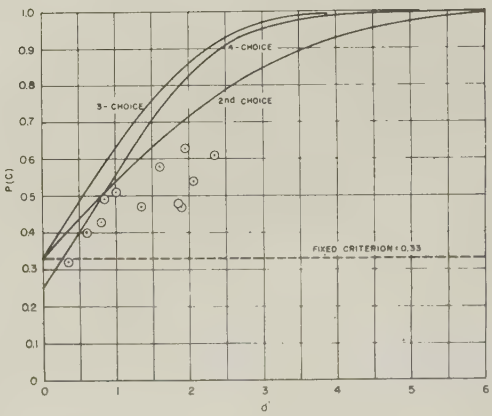


Fig. 14 - Second-choice data.

THE HUMAN USE OF INFORMATION

II. SIGNAL DETECTION FOR THE CASE OF AN UNKNOWN SIGNAL PARAMETER*

Wilson P. Tanner, Jr. and Robert Z. Norman
University of Michigan

Abstract

Two specific cases of signal detection involving uncertainty in the frequency of a sound signal are compared with the case of the signal-known-exactly. In the first case the signal is either of two known frequencies; in the second case the signal is any frequency within a given range. It is suggested that detection behavior that is optimal for the three cases requires a dual mechanism: a combination of a wide-open receiver and a panoramic receiver. Evidence is presented that supports the existence of such a mechanism. Estimates of the bandwidth and scan-rate of the receiver are included.

Introduction

This paper is one of a series in which receiver theory is applied to human sensory behavior. This is a logical application for the human sensory systems are, of course, receivers, picking up transmitted energy and transforming this energy to a useful form.

The observable aspects of the system are (1) the input to the system, and (2) behavioral acts based on an interpretation of the output. Inferential knowledge of the receiver characteristics can be gained by a study of the observable data, with some help from physiological studies of the sensory systems of infra-human animals.

Generally, empirically derived relations between the two sets of observable data have failed to furnish an adequate basis for understanding the sensory systems. A more fruitful approach lies in the construction of a hypothetical model based on simple assumptions consistent with known physiological data. This model must lead to predictions consistent with physiological data, and it must also be capable of generating new hypotheses. It is for these reasons that a model based on the theory of statistical decision (or the theory of testing statistical hypotheses) was selected as appropriate. This model suggests considering the sensory systems as receivers subject to internal noise (an assumption consistent with physiological data); in addition, this model requires the assumption that the output of the sensory systems is treated in an optimum manner (in effect, a new hypothesis).

The first step in the development of the statistical decision model for human sensory behavior was taken by Tanner and Swets.[1] Their results show that for the case of the signal-known-exactly in visual detection the optimization assumption is reasonable. There appears to be a mechanism capable of behaving as an hypotheses-testing mechanism which acts on the basis of likelihood ratios at the output of the visual pathways. Enough of this experiment has been repeated for the auditory case of detecting signals in noise so that, when considered in conjunction with the evidence of Smith and Wilson[2], the model can be considered applicable to audition as well as vision.

Tanner and Swets were concerned with the case of the signal-known-exactly. This paper is concerned chiefly with detection for a case where the signal is not known exactly. The particular case is that in which there is uncertainty in the frequency of a sound signal which appears in a noise background. Two specific cases will be compared with the case of the signal-known-exactly: (1) the signal is at one of two frequencies with the separation of the frequencies as a parameter, and (2) the signal is at any frequency within a given range.

These three simple detection problems point out that care must be exercised in applying the optimization assumption. Each situation, considered alone, requires a somewhat different receiver or combination of receivers for optimum behavior. It thus becomes apparent that one of the criteria for selecting the hypothesized type of optimization is the compatability of the number of different receivers required by the particular type of optimization with present knowledge of neurophysiology. It is unlikely that a separate receiver exists for every possible laboratory situation. It seems necessary, therefore, to try to find a single receiver which is optimum for the three laboratory situations outlined above, or, if more than one receiver is to be called into play, to insist that such additional mechanisms must be capable of being justified on the basis of more general considerations, for example, biological utility. For example, for the three situations of concern in this paper, a

* This paper is based on work done for the U. S. Army Signal Corps under Contract No. DA-36-039 sc-15358.

multiplex receiver that is capable of handling any combination of the three experimental situations would be the optimum receiver. The existence of such a receiver, however, is not compatible with the data presented below. A panoramic receiver, with a scan rate determined by the task, is a receiver near optimum for these three tasks, although this falls considerably below the multiplex receiver for the case where the signal is one of two frequencies. Actually, the latter is not a biologically significant case, and consequently should be given a minor role in the application of the optimization assumption.

A Dual Mechanism

There is a biologically significant case which suggests the possibility of more than one mechanism for the three cases considered above. This is the case where it is optimal for the animal to attend to specific signals, and, at the same time, to be warned in the event that something occurs outside of the range of signals to which he is attending. The attention to a specific signal requires a narrow-band receiver, the warning requires a wide-open receiver. A much over-simplified neural system permitting the operation of such a dual system is illustrated in Figure 1. The columns labled R are the receptors A, B, C, D, E. The columns labeled N are neurons A, B, C, D, E, and W. If a signal is detected at the output of a neuron (A to E) the receptor from which the signal originated is known. If one is detected at the output of W, however, the only information is that a signal exists, originating from at least one of the receptors. The information from W is that provided by a wide-open receiver; from A to E the information is that provided by narrow-band receivers.

Now, if the animal is attending to E, for example, occasional inspection of W may serve to determine the existence of signals other than those originating at E. If one exists, then A to D can be considered individually. This arrangement may be far more efficient than examining A to D periodically for the warning. It may thus be reasonable to look for a dual mechanism, a combination of a wide-open and a panoramic receiver. If this is the case, attention should be divided between W and the center of attention, depending on the a priori probabilities of signals over A to E. If there is no probability of signals other than those to which attention is directed, then W should be ignored. If two signals are sufficiently close together, such that the panoramic receiver (with controlled scan range) offers the better probability of detection, then again W should not be observed. If the two signals are farther apart, then either a wide-open receiver or some combination of a wide-open and a fixed-tuned receiver (panoramic, with zero range) should be called into play.

The Bandwidth Problem

From the above discussion it is apparent that one of the variables relevant to the problem is the bandwidth of the receiver in operation. Several writers have reported data bearing on the bandwidth question. In general, there is good agreement on this subject. The results of these studies are reproduced in Figure 2, taken from Licklider.[3] These studies were all performed under different experimental conditions; all, however, for the case of the signal-known-exactly. The material presented below is in agreement with the data represented in Figure 2 to the extent that bandwidth is regarded as a similar function of frequency. This paper, however, differs with respect to estimates of the width of the band. The estimate of bandwidth represented in Figure 2 depends upon an arbitrary assumption that a signal is just audible when its acoustic power is the same as that of the masking noise; this assumption is not suscribed to here.

Green[4] has recently completed two studies in the Electronic Defense Group laboratory which bear on this problem. The first of these studies involved the comparison of an inferred d' with a calculated ideal based on a 10 cps bandwidth.* The principle involved in the study is illustrated in Figure 3. For low values of signal-to-noise ratio (S/N), d' varies as a power of S/N, and is less than the calculated ideal. As S/N increases, d' rapidly approaches the calculated ideal at a d' of the order of 5. d' cannot, of course, exceed S/N. This suggests that the bandwidth may be as narrow as 10 cps.

Green's second study involves the problem of matching bandwidth to signal duration. For durations less than about .08 second at 1000 cps, d' is a linear function of signal duration, t. For durations greater than about .08 second, d' varies as $\sqrt{t}$. Thus, 12.5 cps appear to be the maximum bandwidth. Now, suppose the observer knows a signal is .2 seconds in duration. If he is still operating with a 12.5 cps bandwidth, signals of .1 second in duration, introduced without his knowledge, should result in a d' which is .707 of the same signal at .2 second. If, however, he has matched his bandwidth, narrowing it to 5 cps, then the signal of .1 second duration, again introduced without

* For the definition of d' see Tanner and Swets.[1]

knowledge, should yield a d' of 1/2 that of the same signal energy .2 second in duration. In an exploratory experiment, the d' for the .1 second signal was observed to be exactly 1/2 that of the .2 second signal. The likelihood ratio comparing the matched bandwidth against the fixed bandwidth was 3:1 in favor of the matched bandwidth, suggestive but scarcely conclusive. The difference between the fixed-bandwidth prediction and the experimental result is significant at the ten percent confidence level. This work is being continued.

In general, Green's work shows that the maximum possible bandwidth is a logarithmically increasing function of frequency. The nature of this function, for the range of frequencies investigated (500 cps - 4000 cps), can be described approximately by the equation

$$\omega = 10^{kf} \qquad (1)$$

where ω is an inferred measure of bandwidth, f is frequency, and k is an individual constant.

The Experiments

The experiments to test the mechanism described above were designed to answer the following questions. 1) Can the hearing mechanism act as a fixed-tuned receiver? 2) When two signals are separated in frequency is it possible to listen for both at the same time? 3) Is the scanning hypothesis feasible? 4) Are there situations in which the data can best be described in terms of a wide-open receiver? The procedures for testing were the same throughout.

Procedure

A forced-choice experimental technique was used. All programming was carried out by N. P. Psytar[5] (Noise Programmed PSYchophysical Testing And Recording). The observers listened with Permoflux PDR-8 cushioned headphones to a signal presented by a tone burst generator in a background of white noise. The signal occurred simultaneously with one of four flashes of a neon bulb, and the observer's task was to state with which of the flashes the signal occurred. Wherever the experiment involved a comparison, such as that between the signal-known-exactly and the signal known to be one of two frequencies, the comparison was based on a single day's data if at all possible.

Experimental Evidence

The Ability to Act as a Fixed-Tuned Receiver. The first, and simplest, experiment merely involves the ability of the observer to tune to a specific frequency to the exclusion of others. The training period, during which the observers became acquainted with the apparatus and became used to listening for signals in noise, was conducted employing only a 1000 cps tone burst .143 second in duration. When they had progressed sufficiently so that no further learning effects were anticipated, the frequency of the tone was switched to 1300 cps, at the same energy level which in the noise background yielded a P(C) (probability of correct choice) of approximately .65 at 1000 cps. The observers were not informed of the change. P(C) for the four observers was approximately chance. They insisted that the experimenter had forgotten to turn on the signal generator. Later tests showed that when they knew the frequency, the P(C) at that signal level, noise level, and frequency (1300 cps) was again approximately .65. It is apparent from this experiment that the hearing mechanism can act as a narrow-band receiver. Unfortunately the nature of the experiment is such that a systematic set of similar experiments (varying the frequency difference between the expected and unexpected signals) is impossible with a single set of observers.

Simultaneous vs. Successive Observation. For the case where the signal is known to be one of two frequencies, different hypotheses lead to different predictions of detection rates. Figure 3 shows the predictions based on simultaneous observation and on successive observation compared to the case of the signal-known-exactly. The curve for simultaneous observation assumes the signals are sufficiently separated in frequency to be clearly resolved by the receiver, while the curve for successive observation assumes a rectilinear passband, such that signals outside of the band are infinitely attenuated. Both curves are probably a little lower than they should be.

In a series of experiments in which the two frequencies were below 2000 cps and were separated by from 200 cps to 800 cps, the results suggest that for all of the separations and for durations of about .05 second, simultaneous observation is impossible. Only in a few individual experiments were the results consistent with simultaneous observation, and these few could have occurred on a chance basis if the successive-observation hypothesis holds.

Evidence for the Scanning Hypothesis. The comparison, however, is a function of the signal duration. If two experiments are run comparing detection for the case of a signal with known frequency versus one of two known frequencies, and all of the parameters of the two experiments are the same except signal duration, then performance with the shorter duration might be expected to suffer more from the lack of knowledge in the two-frequency cases. Such experiments were conducted using frequencies of 400 cps and 1000 cps. The signal durations were .05 and .2 seconds. In each case, for a

single frequency, the signal-noise ratio was adjusted for a P(C) of .8. When the frequency was one of two known values, detection for the .2 second duration was significantly greater than for the .05 second duration. This supports the scanning hypothesis.

Evidence for the Wide-Open Receivers. If the observer knows only that the signal will be in a given frequency range, and this frequency is varied randomly from trial to trial, the probability that the observer is looking for the signal frequency at the time of its occurrence is very low. If the auditory sensory system acts like the narrow-band receiver described in the experiments reported above, P(C) should drop to about .25, the chance probability. Our observations indicate that this is not quite true, although the detection rate does drop well below that for the case of one of two known frequencies. In this case, the hearing mechanism appears to act as a wide-open receiver, not nearly as sensitive as the narrow band receiver because it is open to noise as well as to the signal. It is this case, along with the biological utility of such a mechanism, that leads to the inclusion of a wide-open receiver, necessitating the postulation of a dual mechanism.

An Estimate of Attainable Scan Speeds. The experiments reported above support the hypothesis that the hearing mechanism is indeed a dual mechanism. One part of the mechanism operates as a wide-open receiver, while the other operates as a panoramic receiver. These experiments, however, tell little about the parameters of the panoramic receiver. Apparently it can scan either at 0 speed (fixed tuned as in the case of the initial experiment reported) or at some speed greater than 0. If one is willing to make certain assumptions, it is possible to say a little more about the scan-rate parameter. For example, if a linear scan covering the range determined by the two frequencies is assumed, it is possible to estimate scan rate. While this assumption is probably not realistic, it is made here for the purpose of presenting some preliminary calculations of scan-rate.

Two experiments were involved in this study. In each experiment the signal could occur anywhere within two frequencies: 400-1100 cps and 1000-1700 cps, respectively. P(C) for signals in each range were determined for signals of known frequency of .1 second duration. Then the unknown frequency experiment was done, increasing duration until the known P(C) was again achieved. In the lower frequency range it was necessary to increase the signal to approximately .3 second to achieve this detection level, while in the higher range it was necessary to increase the duration to approximately .2 second. Thus, the high frequency range, which is 700 cps wide, can apparently be scanned at a rate approximately 1.5 times the scan rate in the lower frequency range, which is also 700 cps wide.

Thus, in the low range it is possible to scan over the frequency range (700 cps) in something like .2 second, and at the higher range in approximately .1 second. Assuming linearity over the range only, the scan-rate is 3500 cycles per second per second in the lower frequency range, while in the higher range it is 7000 cycles per second per second. Arbitrarily assuming that these are the rates for 700 cps and 1400 cps (approximately the mid-frequencies of the two ranges) the scan-rate may be approximated by the equation

$$\frac{df}{dt} = 5f. \qquad (2)$$

The rate of change in scan-rate thus appears to be a linear function of frequency.

Conclusions

The hearing mechanism is treated as a dual mechanism and experimental evidence is presented supporting the feasibility of this treatment. The two components of the mechanism are 1) a narrow-band panoramic type receiver, and 2) a wide-open receiver. The employment of these receivers is under control of the individual and dependent upon the type of task he is asked to perform. When frequency information is either available, or required, the narrow-band receiver is used. When one is trying to detect only the presence of a signal, the wide-open receiver is employed.

The experiments performed involved signals at a level seldom significant in real life situations. This was necessary for the purpose of the study. It also leads to some statements that, at first glance, suggest the existence of behavior that is biologically detrimental. For example, the ability to attend to a single frequency to the exclusion of others differing by a relatively few cycles could lead to disastrous events. The fact that the level of signal employed is so low leads to this result. Signals at higher levels either may not be attenuated to so great a degree, or may be sufficient in amplitude for detection with occasional reference to the wide-open receiver, particularly if they are of sufficient duration to be significant to the individual.

Another problem arises, and this is the perception of speech. For speech to be perceived, the panoramic receiver is required. The antics it must perform if it is to follow the sound frequency patterns are scarcely imaginable. The obvious conclusion is that the receiver does not follow these sound patterns exactly. It searches on the basis of conditional probabilities, frequently failing in

the search. When it fails, the undetected frequencies are filled in on the basis of a posteriori probabilities. It is for this reason that a phoneme improperly used or improperly articulated may not be detected, and a more likely phoneme substituted in its place by the listener. When substitutions of this sort are made, the listener is usually convinced that he heard the more likely phoneme. He seems to be unaware of having made a correction.

There is still a great deal of work necessary to complete the picture. Some of this work is in progress, including parallel studies in the visual area to see if the dual mechanism best describes the case where signal location is unknown. The problem is complicated by the possibility of a non-linear scan, and further progress depends on determining the nature of the non-linearity and more precise information on scan-rates.

References

1. Tanner, W. P., Jr., and Swets, J. A., "The Human Use of Information. I. Signal Detection for the Case of the Signal-Known-Exactly." Transactions of the I.R.E. Professional Group on Information Theory (This issue).

2. Smith, M., and Wilson, Edna A., "A Model of the Auditory Threshold and Its Application to the Problem of the Multiple Observer." Psychol. Monog., Vol. 67, No. 9, 1953.

3. Licklider, J. C. R., "Basic Correlates of the Auditory Stimulus," in S. S. Stevens (Ed.), Handbook of Experimental Psychology , New York: Wiley, 1951.

4. Green, D. M., "Signal Detection as a Function of Frequency and Duration," In Technical Report No. 30, Electronic Defense Group, University of Michigan (in preparation).

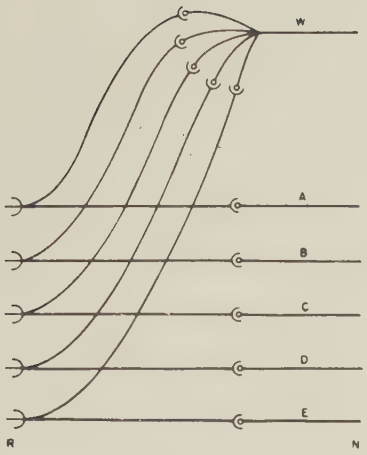


Figure 1
A Simplified Model
of the Hearing Mechanism.

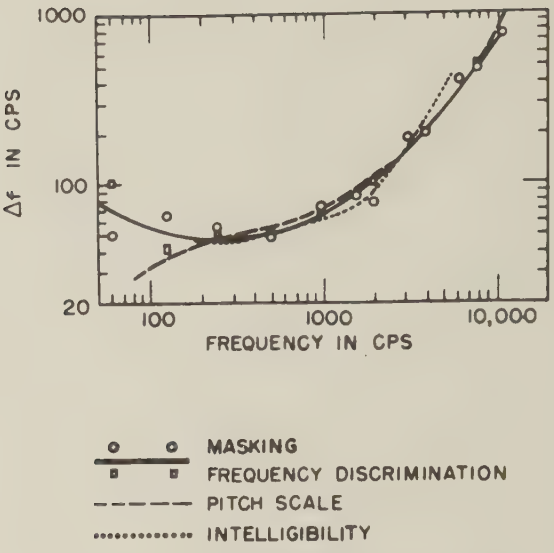


Figure 2
Estimate of Bandwidth as a Function of
Frequency. (Taken from Licklider[3])

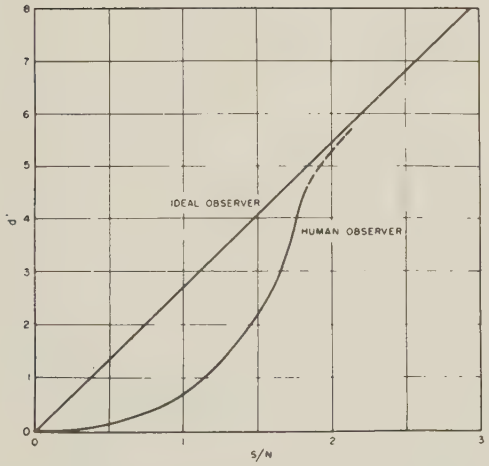


Figure 3
Human Observer Compared to Ideal Observer.

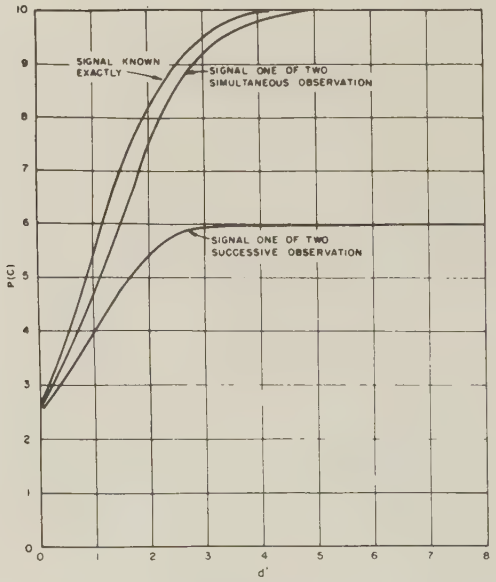


Figure 4
Case of Simultaneous and Successive
Observation Compared to the Case of
The Signal Known Exactly.